THE UNIVERSITY
of EDINBURGH

A Methodology to Sustain Common Information Spaces for Research Collaborations

Luca Trani

Doctor of Philosophy

Centre for Intelligent Systems and their Applications
School of Informatics
University of Edinburgh

2019

Abstract

Information and knowledge sharing collaborations are essential for scientific research and innovation. They provide opportunities to pool expertise and resources. They are required to draw on today's wealth of data to address pressing societal challenges. Establishing effective collaborations depends on the alignment of intellectual and technical capital.

In this thesis we investigate implications and influences of socio-technical aspects of research collaborations to identify methods of facilitating their formation and sustained success. We draw on our experience acquired in an international federated seismological context, and in a large research infrastructure for solid-Earth sciences. We recognise the centrality of the users and propose a strategy to sustain their engagement as actors participating in the collaboration. Our approach promotes and enables their active contribution in the construction and maintenance of Common Information Spaces (CISs). These are shaped by conceptual agreements that are captured and maintained to facilitate mutual understanding and to underpin their collaborative work.

A user-driven approach shapes the evolution of a CIS based on the requirements of the communities involved in the collaboration. Active users' engagement is pursued by partitioning concerns and by targeting their interests. For instance, application domain experts focus on scientific and conceptual aspects; data and information experts address knowledge representation issues; and architects and engineers build the infrastructure that populates the common space.

We introduce a methodology to sustain CIS and a conceptual framework that has its foundations on a set of agreed Core Concepts forming a Canonical Core (CC). A representation of such a CC is also introduced that leverages and promotes reuse of existing standards: EPOS-DCAT-AP.

The application of our methodology shows promising results with a good uptake and adoption by the targeted communities. This encourages us to continue applying and evaluating such a strategy in the future.

Lay Summary

This thesis investigates how to facilitate the establishment and sustainability of information and knowledge exchange in research collaborations. We recognise the importance of research collaborations in modern science as drivers of progress and innovation. They often embrace diverse and heterogeneous profiles and type of actors. Each has their own background and set of skills. To enable effective communication and exchange among participants of the collaboration our goal is to provide them with a common shared context or Common Information Space (CIS).

The definition and construction of such a context is a primary target of our research. This CIS not only serves human communication but it also addresses the requirements of automated methods. Those necessitate that the shared context is represented in specific forms e.g. in order to be consumed by computer programs.

The characteristics and features of the CIS need to fulfil the requirements of the actors of the collaboration e.g. scientists, practitioners, experts and managers. To better understand how such dynamic and complex cooperative systems work in practice, we analysed two specific cases: an international federation of seismological data centres and a large research infrastructure that includes ten disciplines studying the solid-Earth. These exposed issues and challenges that communities face when they attempt to collaborate across disciplines and cultural borders.

We recognised that in order to fulfil their needs it is crucial to address socio-technical challenges jointly. This means that aspects such as organisational and social environment, governance and individual background of scientists, practitioners and managers involved in the collaboration, need to influence the shaping and designing of underpinning technical solutions.

This thesis identifies and characterises such challenges and complexities and then proposes and validates an approach to address the collaborative endeavour effectively. That approach delivers a methodology and supporting tools for CIS which will be applicable for any inter-disciplinary collaboration that needs help with its formation or sustainability.

Acknowledgements

Acknowledging all those who influenced me while producing this thesis is a difficult task. I am extremely grateful to my principal supervisor Malcolm Atkinson. In these years of continued interactions, fruitful discussions and intense exchanges he has been an inspiring mentor, an enthusiastic teacher and a true friend. He guided me throughout this endeavour encouraging, understanding and sustaining me during difficult times. It was a real honour to work with him.

Much gratitude goes to my second supervisor Rosa Filgueira. Her thoughtful and valuable advise provided me with new insides, stimulated ideas and helped me in critical moments. I am very thankful to Aurora Constantin, who kindly offered me her support and her broad HCI knowledge by discussing and by reviewing my approach to perform the evaluations.

I am very grateful to my manager Läslo Evers. He has been fundamental in this long journey. This result was not possible without his convinced endorsement and continued support. I would like to thank my KNMI colleagues, in particular those with whom I worked with more closely on aspects related to my research: Jarek Bieńkowski, Jordi Domingo, Mathijs Koymans, Andrea Pagani and Reinoud Sleeman.

I am thankful to all the ORFEUS-EIDA colleagues and those with whom I worked with in the EPOS, EUDAT and EOSC-hub projects, in particular Daniele Bailo, Massimo Fares, Yann Le Franc, Otto Lange, Rossana Paciello and Javier Quinteros. I would like to thank the participants of the surveys reported in this thesis. Their views and feedback were highly appreciated.

Also, I am very thankful to many bright speakers that I met at conferences and meetings; to colleagues with whom I exchanged views and ideas; to several people with whom I engaged in stimulating discussions; they provoked useful thinking and were valuable inspirations and motivations.

A special mention goes to Torild van Eck, his memory and intellectual heritage are still vivid and keep inspiring those who had the pleasure to work with him.

Finally, I would like to thank my beloved family, my partner Lucinda and my two kids Giovanni and Simon. During these years they have been my greatest source of energy and joy.

Declaration

I declare that this thesis has been composed by myself and that this work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. These cases are enumerated here and in the Publications section 1.6. Use cases and requirements supported by the approach described in this thesis were discussed and validated within a team of domain and technical experts in the context of the funded H2020 project EPOS-IP and the ORFEUS-EIDA federation. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others. Contributions and material provided by other authors to this work have been explicitly indicated below.

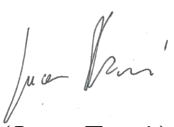
Chapter 4. Mathijs Koymans, a KNMI colleague, helped under my supervision with the implementation of the `WFCatalog` system and the quality tools. He kindly provided figure 4.3.

Chapter 5. Figure 5.3 was kindly provided by the EPOS Management Office. Figure 5.4 draws on data collected by EPOS WP6-WP7 colleagues.

Chapter 6. Mathijs Koymans helped with the collection of `WFCatalog` usage statistics reported in Table 6.1. The questionnaires described in Section 6.2 were submitted during an EPOS meeting where Daniele Bailo and Rossana Paciello helped organising the surveys.

Appendix A. The EPOS-DCAT-AP model in Figure A.1 was initially presented to and discussed with colleagues in the EPOS metadata Task Force. It was then made available on a public GitHub repository where it received comments and feedback from other EPOS participants and data experts.

Appendix B. Rossana Paciello helped by completing some of the SHACL shapes graphs reported in Listing B.1. Other colleagues, as indicated in the listing, contributed their feedback.



(*Luca Trani*)

A mio padre

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Science benefits tremendously from mutual exchanges of information and pooling of knowledge, effort and resources. The combination of different skills and diverse expertise is a powerful capability, source of new intuitions and creative insights. Therefore multidisciplinary, holistic approaches can be a great opportunity to explore novel scientific horizons. Collaboration is not only an opportunity, it is essential when tackling today's global challenges by exploiting our fast growing wealth of data.

This thesis delves into issues and socio-technical challenges associated with the establishment of effective collaborative practices in research infrastructures. To capture the requirements and opportunities of such scientific endeavours an abstraction is introduced – the concept of Information-Powered Collaborations (IPC). We investigate the inherent complexity associated with such dynamic environments and propose an approach that partitions such complexity and offers concrete tools and methods to thrive in the data revolution era.

1.1 Collaboration and data sharing

Cooperation and collaboration have characterised the organisation of work in various contexts throughout history. Consequently, the support for collaborative work has been investigated for a long time by scientific disciplines such as the Computer Supported Cooperative Work (CSCW). Since the mid 80s a rich CSCW literature produced several theories and approaches proposed to model and improve collaborative work sustaining sharing of knowledge and expertise [Star and Griesemer, 1989; Ackerman

et al., 2002, 2013]. The importance of scientific collaborations is not only well-recognised but it is encouraged and fostered, *e.g.* by policy makers and funding bodies, as a way to improve impact, to achieve cost-efficiency and to tackle the pressing data challenges faced by nearly all scientific disciplines. Collaborations based on information sharing, contribute different viewpoints and combine skills and intellectual efforts to tackle the increasing complexity of contemporary scientific challenges. Establishing effective collaborations among diverse actors involves inherent socio-technical issues and necessitates "*a common terminology and shared knowledge base*" [Lubich, 1995a].

In modern (data-driven) sciences data sharing is a key to enable successful collaborative behaviours. As such, it has received considerable attention in the last decade being widely recognised as an accelerating factor for the scientific progress [Fecher et al., 2015].

Atkinson et al. introduce the concept of '*Data-to-Knowledge highways*' [Atkinson and Parsons, 2013] where they highlight the importance of combining multiple sources of knowledge. Building such highways requires the strong collaboration of several skills and expertise, identified by the authors with three categories of expert: *Domain experts*, *Data-analysis experts* and *Data-intensive engineers*. By working closely together these actors, or roles, deliver tools and methods to extract information from the increasing wealth of data and provide it in the forms required to drive informed decisions.

Data sharing is just one aspect underpinning research collaborations. Equally important are: sharing of methods, context and best practices; understanding of implicit communication rules, norms and prior knowledge that form the culture of the involved scientific communities (*designated communities*) [CCSDS, 2012]. Many aspects of the culture, such as formalised methods and data-access rules may be represented as shareable data so that extensive distributed collaboration can be better supported.

Building research collaborations is a major endeavour that requires time and investments that increase rapidly with the diversity and the number of involved parties. Retaining the value of those investments, sustaining and maintaining efforts over time are necessary strategic choices. The management of research collaborations ought to interface and account for the organisational structures present in each community.

Different strategies may be needed to address these issues – they are investigated in this thesis.

A key goal is to establish effective communication channels in order to convey the correct information to the parties involved in the collaborations and stimulate their mutual understanding. Such channels ought to account for a variety of requirements.

Even a single scientific discipline might reveal heterogeneity that varies depending on aspects such as domain, focus, maturity and objectives. In some cases there are well-defined practices and widely adopted standards, thus communication and data exchange flow following well-known behaviours and patterns. Typically, well-established communities with a long history of close collaboration fall in this category. Another interesting category, which is rapidly developing and gaining relevance, is constituted by those disciplines belonging to the so called 'long-tail' of science [Borgman et al., 2015]. This group includes a variety of scientists belonging to different disciplines who generate, manage and share relatively small amounts of highly variable and heterogeneous data. Typically, they develop hand-crafted, *ad hoc* data management solutions which make data and information exchange difficult.

To model the complexity of those sharing contexts we introduce the concept of **Information-Powered Collaborations** (IPC) below.

Definition 1. *Information-Powered Collaborations (IPC): are complex, dynamic and heterogeneous environments that enable information sharing among actors (e.g. researchers, scientists, practitioners, agents) from independently managed organisations (e.g. research institutes, resource providers), thereby supporting knowledge and expertise exchange in a multidisciplinary context. The resulting collective knowledge can be harnessed to accomplish common scientific goals and foster novel scientific viewpoints drawing on the integration of the diverse perspectives.*

IPC is an abstraction that we use extensively in this thesis and represents a typical modern research context characterised by rich interactions, exchanges and complex dynamics. Traditionally the research scene was dominated by research groups in controlled environments, with limited interactions with their peers [Lubich, 1995b]. The data revolution has deeply impacted every domain demanding a paradigm shift where collaboration is essential to manage the amount of data and to interpret the derived information. The IPC can offer a means to address and tackle today's

challenges by stimulating and by facilitating *pooling of knowledge and expertise* (as well as data and information). In the next section we introduce some aspects that help us characterise and describe our research focus.

1.2 Problem definition

IPC constitute the landscape in which this research is scoped. They provide us with a wide and challenging context, which because of its complexity cannot be addressed entirely in one PhD. Whilst keeping in mind a long-term vision and a broader view, our analysis focuses on specific issues exposed by such multifaceted digital ecosystems. A number of aspects are identified, discussed and addressed in this thesis. In our view those are essential components to establish sustainable processes that support pooling of knowledge and stimulate innovation in IPC.

For instance, we recognise the importance of defining, representing and populating common shared contexts, identified with the CSCW concept of *Common Information Spaces* (CISs) [Bannon and Bødker, 1997] – they are exploited to drive communication, exchange, interoperability and innovation. We acknowledge the centrality of the role of the communities participating in the collaborations, *designated communities* – they must *retain control* over the construction, maintenance and evolution of their CISs.

In the following subsections we introduce relevant concepts that support our problem definition.

1.2.1 Terminology and context

People communicate leveraging backgrounds established by heritage, culture, education, experience, *etc.* Typically members of one community use concepts and their associated terms from their backgrounds implicitly. As they share terminologies and interpretations of such concepts acquired in a common context, communication and exchange are facilitated. Details about specialisations and customisations might still be argued and yield disputes (*e.g.* synonymy, homonymy, polisemy), nevertheless the common shared space is an accelerating factor for effective interactions.

Terminology and context are deeply interconnected as the latter influences the interpretation of the first; *e.g.* a broader or narrower context for a specific term can

lead to different interpretations. For instance, in a general context a term such as '*root*' has several meanings: part of a plant, or part of a hair, or part of a tooth, or the source of something, or the root of a number, *etc.* Even slight variations such as its plural (*i.e.* roots) carry a different semantics *e.g.* origins of a person. By specifying a context (*e.g.* botany, mathematics) one can reduce the possible choices and clarify the meaning. Hence the importance of context.

People that speak the same language share a common set of grammar rules and a set of terms with associated meanings that can be mutually understood. Terms and their related meanings belonging to a shared context can be understood without explicitly referring to the originating context, thus enabling fluent communication. Conversely, when the actors do not share a common context terms ought to be accompanied by descriptions of their context in order to be understood correctly, avoiding misinterpretations and conveying the right information. The size and content of the additional description required depends on the heterogeneity of origin and target contexts. Excessive and redundant explanation becomes clutter that inhibits communication.

We can think of such a context as a *Universe of Discourse* (UoD) that enables effective communication [Atkinson et al., 2018]. UoDs are dynamic and evolve overtime. For instance, a new UoD can be formed by merging (parts of) existing UoDs to enable the communication in a newly established collaboration involving different communities. UoD is an abstract entity that can be used to model concepts, practices, methods, norms, agreements and rules that form the culture and identity of a specific community. UoDs can be implicitly or explicitly derived, transferred and understood by humans. However, in computer interactions UoDs must be *explicitly* represented and effectively conveyed. It is important that the humans have a consistent interpretation of each term for communication with each other and across time. We might call this their intended meaning. Then computers in their interpretation of that meaning should reflect the human intention. Discrepancies between these interpretations lead to less usable systems and to errors.

Computer systems can enable semantic interoperability and knowledge exchange by sharing such representations and by providing harmonisation and reconciliation mechanisms. Our work targets this conceptual space and researches effective ways to build, represent and maintain a common UoD, thereby enabling pooling of knowledge

and achieving semantic interoperability in multidisciplinary collaborations spanning independent organisations. In this thesis we refer to such a common UoD as a *Canonical Core* (CC).

1.2.2 Catalogues and UoD representation

The variety of concepts and categories forming a UoD can be partially described, formalised and represented in catalogues. Catalogues can be populated with and provide access to UoD's entities and their relationships. They can be valuable tools to form, maintain and sustain formal agreements and to address communities' requirements, we present a selection of requirements in the next section (1.2.3). Metadata are important ingredients to enable machine-readable representations, therefore metadata catalogues can be seen as a means to represent UoDs and convey their information effectively. In this thesis we focus on metadata catalogues that hold representations of Canonical Cores.

1.2.3 Examples of requirements

We identified a number of requirements recurring in IPC – a representative selection is presented in this section.

1.2.3.1 Resource discovery

Resource discovery – implies the search of high-level descriptions (metadata) carrying information for instance, about type, name and origin of a resource. It entails operations such as selection and filtering matching specified criteria. Examples of a multi-faceted search crossing domains: `FIND` all time series catalogued since `date, time` giving geochemical emission, seismic activity and surface movement for Etna; or `FIND` the seismic events in 2017 in Southern Europe together with geology, Global Navigation Satellite System (GNSS) velocity and satellite data that could be correlated with those events.

1.2.3.2 Resource evaluation

Resource evaluation – requires deeper descriptions of resources (domain-specific metadata). It exploits additional metadata fields beyond the classification of a resource in order to query, select, filter actual instances of resources according to desired characteristics. Example: `FIND` all the seismic events with magnitude $M > 5$, that occurred in a time-window `(Tw)`, in a specific region `(Re) AND` the related primary data (seismic waveforms) with fewer gaps than 5% in `Tw AND` the GPS displacement maps associated with `(Tw, Re)`.

1.2.3.3 Scientific methods

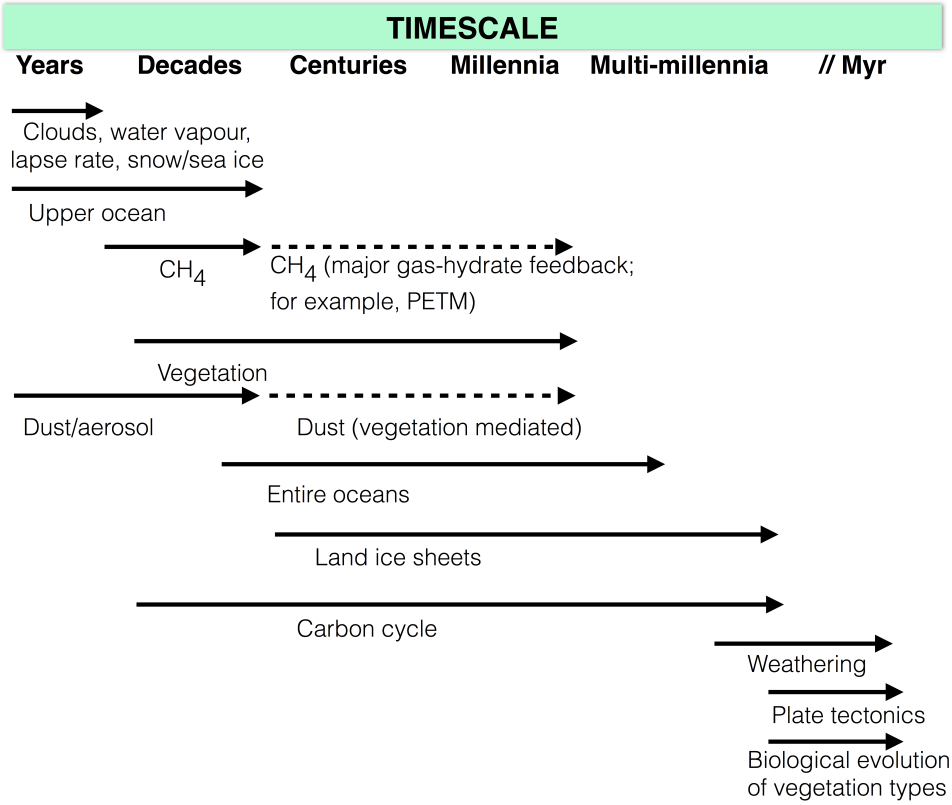
Scientific methods support – helps collaborating teams of experts create and refine methods that draw on the diverse resources and data collections. It promotes the formalisation and automation of these methods, typically as scientific workflows [Atkinson et al., 2017], while supporting critical procedures to deliver good quality evidence contributing to the shared knowledge.

Example: develop methods and models to reveal the impact on seismic hazard from mineral extraction methods. The authoring system consults the catalogue to help the method developer make choices, detect defects and plan enactment. The enactment system consults the catalogue to verify compliance with policies, to plan the optimal deployment and annotate provenance records. The provenance system links with the catalogue, mainly via identifiers, to support diagnostics, validation, reproducibility and evidence qualification.

1.2.4 Supporting shared agreements

Achieving shared agreements on the interpretation of terms in multidisciplinary environments is not a trivial task. Even a common concept such as `time` can carry diverse semantics depending on the temporal reference or the calendar used in specific contexts. For instance, in archeology or geology time is often expressed counting years backwards from a reference date. In a lunisolar calendar (*e.g.* Chinese Calendar) time is expressed according to astronomical phenomena. Those reasons inspired domain-specific formalisations, *e.g.* for geological timescales [Cox and Richard, 2015], and extensions in conventional representations such as OWL-Time [Cox and Little, 2017]

to include non-Gregorian calendars [Cox, 2016]. Figure 1.1 provides an example of the diversity in time scales which are present in solid-Earth sciences. Each of those might be associated with different reference systems and therefore a different semantics of time. The deployed instruments may resolve time with sub-microsecond resolution to triangulate signal sources. A conceptual framework from geological, through historical to observational time needs clarity about the transitions and correspondences.



Source: [Members, PALAEOSENS Project, 2012]

Figure 1.1: Diversity in time-scale nomenclature and precision experienced in research that engages with the distant past as well as the present, such as the solid-Earth sciences. This is just part of the range encountered by sciences that observe to sub-microsecond resolution for today's observations to resolve hypothesised models spanning billions of years.

In order to support multi-disciplinary, multi-organisational and multi-national collaboration the underlying concepts must be recognised and agreed. These are often formalised as ontologies [Marshal, 2011; McGibbney, 2018]. Collaborative development

of such ontologies often reveals variations and encourages refinement of such concepts, illustrating the kind and scale of investment needed to build, agree and adopt a Canonical Core.

1.2.5 Sustainable framework

The time-scale and scope associated with IPC carry important implications. Typically, IPC cover long periods. For instance, the European Centre for Medium-Range Weather Forecasts (ECMWF[1]) is a large international collaboration established in 1975 "*to pool Europe's meteorological resources to produce accurate climate data and medium-range forecasts*" and is currently leading advancements in Numerical Weather Prediction (NWP). The Square Kilometre Array (SKA[2]), an impressive scientific collaborative endeavour to build the world's largest radio telescope, is planned to be operational for over 50 years after its construction.

As more communities engage and endorse an underpinning CC, additional challenges arise to meet and fulfil their requirements. For instance, long-running campaigns spanning tens of years would witness considerable technological and organisational changes. They might hit current limitations and push for revisions in order to incorporate those changes. They will require innovation to make progress towards their targets. At the same time the strong community engagement and endorsement of the CC would promote caution with modifications as most practitioners do not want their established working practices disrupted. Indeed they may be studying trends and find changes reducing the validity of their evidence. The introduction of necessary improvements requires clear paths to encourage take up of new capabilities. Whilst adjustments are inevitable, trustworthiness and reliability should not be undermined.

The value of a framework that supports an IPC and provides a holistic view of the diverse autonomous resources is evident as it supports repeated reuse. However, this implies that it needs to be sustained for the long-term. For the success of such a system it is vital to balance specified structure against agility, stability against change.

[1] www.ecmwf.int

[2] www.skatelescope.org

Figure 1.2: Finding the right balance between well-defined, constrained structure and dynamic, agile structure

1.3 Research objectives

The IPC challenges, introduced in the previous section (1.2), are the driving forces that motivate this research. They converged into two compelling goals and their corresponding sub-goals that are listed below:

G_1 Establish a methodology that enables experts to empower research collaborations by pooling knowledge, expertise and resources from diverse actors.

- **$G_{1.1}$** Deliver strategies for coupling of different viewpoints to enable advances in science.
- **$G_{1.2}$** Motivate the continued and effective engagement of the actors in the collaborations.

G_2 Develop a framework for creating and sustaining holistic views of diverse, evolving, independent information resources.

- **$G_{2.1}$** The framework should have the potential to cope well with realistic diversity and dynamics encountered in anticipated research federations.
- **$G_{2.2}$** The benefits from the framework should outweigh the complexity costs so that future adoption and maintenance will be feasible.

In the scope of this thesis we developed and pioneered a methodology that addresses significant parts of both goals (*i.e.* G_1 and G_2). Because of the breadth of our target we validated of our approach heuristically, and collected relevant and promising

indications about its validity and feasibility. However, at present we are not able to demonstrate the pursued benefits in the long-term. To address those we propose an approach to perform continued measurements that can be harnessed to monitor progress. Long-term results will depend on human behaviours, socio-technical aspects, governance issues, *etc.*, thus systematic assessments will be required to capture them. In the next section we describe the strategy adopted in this research and summarise the results achieved.

1.4 Research contributions

We observed how research collaborations depend on bridging boundaries and sharing information from autonomous sources. The concept of Information-Powered Collaborations (IPC) was introduced as an abstraction to characterise those rich environments and their complex interactions. In our research we analysed prominent examples of IPC as they were developed in the context of large research infrastructures such as: the European Plate Observing System (EPOS[3]) and the Observatories & Research Facilities for European Seismology (ORFEUS[4]). Engaging in these initiatives provided us with inspiration, concrete requirements, use cases and scenarios coming from heterogeneous scientific communities with a particular focus on the solid-Earth sciences. Those large federated infrastructures constituted a solid platform and a '*test range*' where we developed and applied the approach described in this thesis. Such an approach was derived by actively participating in those research collaborations and by observing their interactions and dynamics.

We started from a well-governed seismological environment, *i.e.* ORFEUS, by addressing the needs to form a shared information space with a defined focus (*i.e.* on the quality of seismic waveform data). In that context we obtained valuable results and learned important lessons that shaped our further investigations. For instance, we understood the centrality of human interactions that lead to necessary conceptual agreements and the influence of socio-technical challenges that had to be addressed jointly. There we started devising a methodology to stimulate and maintain engagement of the involved actors. Figure 1.3 illustrates the CRP methodology that is described in

[3] www.epos-ip.org

[4] www.orfeus-eu.org

this thesis. CRP entails three dimensions to be addressed independently: *Conceptual definition* (C), *Representation* (R), *Population* (P). C develops the ways of thinking, introducing terminology and meaning into the knowledge space, R develops detail and how to organise concepts, P deals with the gathering of instances of concepts in order to meet a community's requirements.

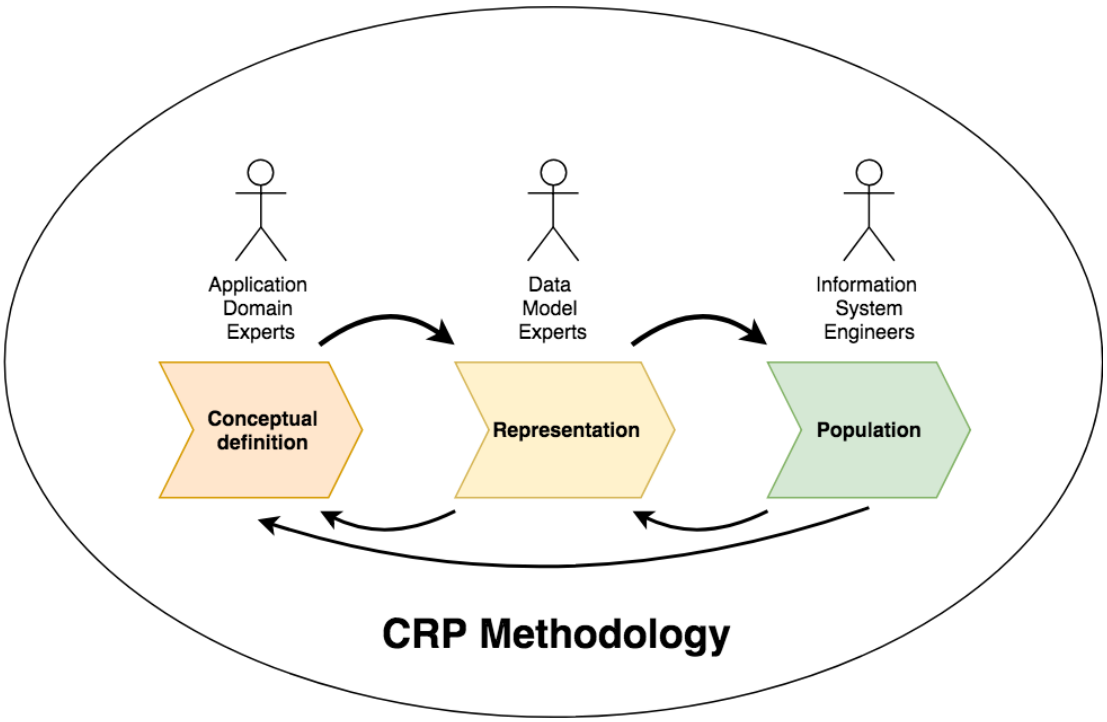


Figure 1.3: Overview of the CRP Methodology

Such a methodology should support exchange, understanding and communication among individuals and deliver solutions for data and method integration, thereby facilitating the pooling of knowledge, resources and expertise. It should have an understandable and recordable representation supported by tools.

In the next phase we targeted a more challenging example of IPC, *i.e.* EPOS. That large and heterogeneous research infrastructure offered us a much wider range of campaign goals undertaken by a diverse cluster of communities that have been brought together recently. It included ten different scientific disciplines with their associated communities. We refined and applied our methodology to address the goals described in Section 1.3 in the EPOS context. We investigated the construction and application of Common Information Spaces as underpinning shared contexts for IPC. They can

be partially represented in canonical forms and maintained in centrally managed catalogues. In this thesis we advocate a layered approach to build such CIS based on a stable Canonical Core, populated with agreed Core Concepts, connected with dynamic Boundary Regions. We identified key aspects of the Canonical Core and proposed a methodology that provides abstractions over the diverse independent sources, partitioning the information space into manageable units. We described the conceptual space, the implications and requirements – organisational and technical – encountered when setting up a such a methodology. Logical structures and representations were explored and designed by relating them to emerging standards that were influenced by our approach. We shaped tools to facilitate their adoption.

Finally we evaluated impact and uptake of our solutions in both contexts, ORFEUS and EPOS, and assessed their influences in the timeframe available for this research. This delivered the contributions described in this thesis that are summarised below:

C_1 An analysis and description of the socio-technical challenges for setting up and sustaining CIS that underpin Information-Powered Collaborations.

C_2 WFCatalog – a concrete and widely adopted implementation that includes a representation of an agreed and shared canonical form for seismological waveform data.

C_3 A methodology to characterise and partition requirements and challenges of IPC. It draws on a Canonical Core formed by agreed Core Concepts and tackles the related issues progressively by exploiting a separation of concerns.

C_4 An architectural framework for a Common Information Space in an IPC that exploits catalogues to represent an agreed Canonical Core and an exploration of its external relationships and evolutionary procedures.

C_5 EPOS-DCAT-AP – a data model to represent a first variation of a Canonical Core that is adopted and developed in EPOS.

C_6 An assessment of the usability of the proposed methodology and a conceptual framework to perform successive measurements.

These are relevant and important achievements, as showed by the recognition and uptake of our work by the target communities. Nevertheless, the complexity and

ambition of our goals will require future investigations as we outline in the conclusions of this thesis.

1.5 Thesis structure

The reminder of this thesis is organised as follows.
Chapter 2 explores conceptual foundations for information sharing. It reviews methods to support collaborative work and platforms for data sharing.
Chapter 3 presents a review of approaches for representing and populating Common Information Spaces. Methods, tools and frameworks to build, manage and exchange populations of concepts are explored.
Chapter 4 introduces the challenges of establishing shared information encountered in a seismological context leading to the adoption of `WFCatalog` by ORFEUS.
Chapter 5 introduces our methodology and describes its application in the EPOS context. It illustrates details of the dimensions of a Canonical Core, their requirements and implications.
Chapter 6 provides evaluations of the results of application of our approach. It defines assessment criteria and proposes a framework for repeated measurements.
Chapter 7 presents conclusions and future work.

1.6 Publications

In the course of this research we produced a number publications, a selection of which is listed below in temporal order. They include journal papers, conference papers and abstracts that provided useful material for this thesis. Some of them (as indicated later) are reported as parts of the chapters, others contributed to shaping thoughts and forming ideas.

- **Trani, L.**, Koymans, M., Sleeman, R. (2016). *Efficient discovery and access to seismological waveform data in ORFEUS EIDA*. Presented at the 35th General Assembly of the European Seismological Commission. URL: `http://meetingorganizer.copernicus.org/ESC2016/ESC2016-335.pdf`

- Bailo, D., Ulbricht, D., Nayembil, M.L., **Trani, L.**, Spinuso, A., Jeffery, K.G. (2017). *Mapping Solid Earth Data and Research Infrastructures to CERIF*. Procedia Comput. Sci. 106, 112–121. doi:10.1016/j.procs.2017.03.043

- **Trani, L.**, Koymans, M., Quinteros, J., Heinloo, A., Euchner, F., Strollo, A., Sleeman, R., Clinton, J., Stammler, K., Danecek, P., Pedersen, H., Ionescu, C., Pinar, A., and Evangelidis, C. (2017). *The European seismological waveform framework EIDA*. In Geophysical Research Abstracts, volume 19. URL: `https://meetingorganizer.copernicus.org/EGU2017/EGU2017-13770.pdf`

- **Trani, L.**, Koymans, M., Atkinson, M., Sleeman, R., Filgueira, R. (2017). *WFCatalog : A catalogue for seismological waveform data*. Comput. Geosci. 106, 101–108. doi:10.1016/j.cageo.2017.06.008

- **Trani, L.**, Atkinson, M., Bailo, D., Paciello, R., Filgueira, R. (2018). *Establishing Core Concepts for Information-Powered Collaborations*. Futur. Gener. Comput. Syst. 89, 421–437. doi:10.1016/j.future.2018.07.005

- Pagani, G.A., **Trani, L.** (2018). *Data cube and cloud resources as platform for seamless geospatial computation*. Proc. 15th ACM Int. Conf. Comput. Front. - CF '18 293–298. doi:10.1145/3203217.3205861

- **Trani, L.**, Paciello, R., Sbarra, M., Ulbricht, D., and the EPOS IT Team. (2018). *Representing Core Concepts for solid-Earth sciences with DCAT – the EPOS-DCAT Application Profile*. In Geophysical Research Abstracts, volume 20. URL: `https://meetingorganizer.copernicus.org/EGU2018/EGU2018-9797.pdf`

- Koymans, M., Fares, M., **Trani, L.**, Quinteros, J., and Nagoe, C. (2018). *FAIRYTALE – Towards FAIR Seismological Data Management in the European Integrated Data Archive (EIDA)*. Presented at the 36th European Seismological Commission, Malta.

- **Trani, L.**, Paciello, R., Bailo, D., and Vinciarelli, V. (2018). *EPOS-DCAT-AP: a DCAT Application Profile for solid-Earth sciences*. In 2018 Fall Meeting AGU. Abstract IN31B-33.

Chapter 2

Conceptual foundations for information sharing

This chapter focuses on aspects that target the conceptual dimension (C) introduced in Chapter 1. We report relevant research and provide an overview of the state of the art of methods to support collaborative work and organisational platforms that sustain data sharing behaviours. Substantial contributions in those areas have been produced by the Computer Supported Cooperative Work research which is introduced in the following sections.

2.1 Computer Supported Cooperative Work and knowledge management

In Chapter 1 we presented the importance of scientific collaborations and introduced the Computer Supported Cooperative Work (CSCW) – a research field founded by Irene Greif [Greif, 1988] and "*focused on the role of the computer in group work*".

CSCW investigated the social aspects of knowledge sharing and the systems to support it. Such investigations yielded approaches to define and maintain '*Common Information Spaces*', to represent knowledge for instance by adopting a 'repository model' and/or exchange it via knowledge artifacts and 'boundary objects' [Bannon and Kuutti, 1996; Bannon and Bødker, 1997; Star and Griesemer, 1989; Star, 2010; Ackerman et al., 2002, 2013]. A branch of CSCW research focused on providing

access to and exchanging expertise, recognising the importance of communication and of helping establish connections among 'knowledgeable actors'. For these reasons CSCW research provided a fertile ground for a number of technical solutions currently adopted in knowledge management and collaborative systems. In the next sections we present details of relevant solutions and discuss their implications for our research.

2.1.1 Sharing knowledge

Knowledge sharing is a foundation for successful collaboration and a major topic of this thesis. It can contribute to exchange viewpoints and perspectives about facts, *e.g.* natural phenomena, which then result in a better understanding of them. For instance, in building a climate change scenario the combination of scientific and economic information is essential to provide a comprehensive impact assessment. Sharing knowledge often yields novel discoveries, for instance, recent studies evaluated the influence of variations of Earth's rotation on processes in geo-dynamics by combining different types of observations, *e.g.* seismic and GPS [Levin et al., 2017; Bilham and Bendick, 2017].

To enable sharing of knowledge an important focus of CSCW has investigated the externalisation of information and knowledge and their representation as artifacs and objects – 'objectivation of knowledge' [Star and Griesemer, 1989; Star, 2010; Ackerman et al., 2013]. The main idea motivating this body of research was that the context, culture and social background, or in other words the knowledge of individuals and organisations, could be collected, represented, maintained and shared for present and future use – "*knowledge has to be both 'past' facing and 'forward' facing*" [Krogh and Petersen, 2010]. Before tackling this demanding challenge diverse approaches had focused on the collection of information objects rather than on the exchange and understanding of knowledge. Those approaches yielded a 'repository model' that aimed at building an 'organisational and collective memory' but its application in real systems soon proved to be ineffective and even utopian [Ackerman et al., 2013]. As Bannon and Kuutti pointed out "*information does not simply exist 'out there', but is produced by specific people in specific contexts for specific purposes. While this does not imply that it is bound solely to that whole context, it does mean that one cannot in any straightforward way extract and abstract from this web of signification items of*

'information' which can be stored in some central resource for later use"[Bannon and Kuutti, 1996].

This recognition of the central role of the actors with their background was a major advance. It was built on a perspective introduced by Bannon and Schmidt in 1989 that targeted fundamental aspects of cooperative work such as interpretation and agreed meanings of information. They conceived the concept of 'shared information space' successively refined in the most known 'Common Information Space' (CIS) [Bannon and Schmidt, 1989; Schmidt and Bannon, 1992] – "*A common information space encompasses the artefacts that are accessible to a cooperative ensemble as well as the meaning attributed to these artefacts by the actors [...] Here the focus is on how people in a distributed setting can work cooperatively in a common information space — i.e. by maintaining a central archive of organizational information with some level of 'shared' agreement as to the meaning of this information (locally constructed), despite the marked differences concerning the origins and context of these information items. The space is constituted and maintained by different actors employing different conceptualizations and multiple decision making strategies, supported by technology*" [Schmidt and Bannon, 1992].

A CIS it is not just a repository of information that can be built once and for all, it is a dynamic entity that evolves. "*Cooperative work is not facilitated simply by the provision of a shared database, but requires the active construction by the participants of a common information space where the meanings of the shared objects are debated and resolved, at least locally and temporarily*" [Schmidt and Bannon, 1992]. This vision sheds a different light also on the concept of 'articulation work' introduced by Strauss and well known in CSCW – "*a kind of supra-type of work in any division of labor, done by the various actors*" [Strauss, 1985]. Articulation can be seen as division of labor but also as a way to form agreements on the meanings associated with information in CIS [Bannon and Bødker, 1997]. As collaboration extends across multiple groups of actors articulation work might be necessary to reconcile the meanings of different CIS.

Bannon and Bødker investigate implications of constructing, using and maintaining CIS. They focus particularly on the 'dialectical' nature of CIS – "*CIS are both open and closed*" – and emphasise the interpretative component of such spaces "*the meaning of the terms or objects are not simply 'given', but require an effort of interpretation on*

the part of the human actors who inhabit this space" [Bannon and Bødker, 1997]. To assume a common shared vision actors of CIS ought to feel relatively free to populate that space with objects of their concern. However, to maintain shared meanings and warrant future use of the information contained in such spaces, CIS ought to be sufficiently closed. Hence, the dialectical nature that ties in with the concept of boundary objects – objects that exist in multiple contexts and are "*both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual site-use*" [Star and Griesemer, 1989]. In a way a CIS can be seen as a boundary object "*packaged and being turned into immutables to allow for sharing across contexts and communities*" [Bannon and Bødker, 1997].

It is quite evident that major challenges in CIS are associated with human behaviours. A key issue is to develop and sustain engagement and motivate the actors and contributors of such systems. For a CIS to build and represent a view as complete as possible of a community's culture and knowledge the active participation of actors and stakeholders is essential. Developing and maintaining engagement ought to overcome barriers that inhibit good sharing behaviours and to promote incentives [Borgman et al., 2015; Kim and Stanton, 2013] – "*if knowledge sharing is not rewarded, employees have no incentive to engage in it*" [Ackerman et al., 2013]. Ackerman et al. identified the following categories of issues: motivation, context in reuse, assessments of reliability and authoritativeness, organisational politics, maintenance, and reification. When trying to represent knowledge it is inevitable to provide a subjective view which brings in some contextual information and leaves out aspects potentially relevant for others. This filtering process is mostly inherent and implicit but also partially explicit and required for practical reasons *e.g.* to reduce unnecessary details and to keep the size of knowledge artifacts manageable. Actors who want to (re-)use a knowledge artifact developed by others need to reconstruct the initial context where it was conceived in order to understand it. They might need additional information that was implicitly or explicitly left out. This contextualisation/recontextualisation process has been broadly discussed in CSCW literature with key papers regarding recontextualisation as a "*situated, social action*". Another challenge regards the assessment of reliability and authoritativeness of the information collected and maintained. CSCW has investigated

strategies to assess quality of the information and to enable trustworthiness. The organisational and political contexts certainly influence CIS. For instance, the processes underpinning the construction and maintenance might need to comply with policies and adhere to guidelines established by an authority. Examples of authorities for the seismological domain are reported in Chapter 4 (*e.g.* ORFEUS and FDSN). We will discuss those issues in more detail in Chapter 5. The concept of CIS is very powerful and still central in modern research. The CSCW literature contains several examples of its successful application in different contexts. CIS "*are in some cases constituted for people that are co-present in time and space, whereas in other situations they are distributed across time and space boundaries*" [Bannon and Bødker, 1997].

In this thesis we assume the concept of CIS and focus on the latter (distributed) case. Typically CSCW analysed focused applications with a controlled scale. For instance, a traffic control room or an emergency medical unit can be seen as examples of CIS [Schmidt and Bannon, 1992; Zhang, 2016]. Also, CSCW is primarily concerned with enabling understanding among humans. In our analysis we target software as well. This aspect has several implications that are discussed further in this thesis. We build on CIS and apply such a concept in the challenging context of IPC. We identify and analyse those aspects that are necessary to address the requirements of federations of loosely coupled actors that dynamically form around a specific focus. Groups of individuals and organisations might decide to establish collaborative work that drives their cooperation. However, the related activities might not necessarily be the primary focus of each of the involved actors. Therefore, engagement and participation in an underpinning CIS should be gained by stimulating their interest or by providing incentives. Active participation ought to be earned for instance, by promoting evidence of benefits and advances in the practices of their concern. They ought to believe that the investments done for the 'common' part of a CIS are worthwhile. We will articulate these challenges and related issues in course of this thesis.

2.1.2 Sharing expertise

Another important focus of this thesis targeted by a branch of CSCW research is the sharing of expertise. Ackerman et al. distinguish those studies as 'second generation' where the emphasis is on "*interpersonal communications of knowledgeable actors*

over externalizations in (IT) artifacts" (studies over the latter are defined as first generation) [Ackerman et al., 2013]. There a particular attention is given to the sharing of tacit knowledge that is typically difficult to formalise and embed in an artifact, in order to make it explicit. A more effective way to acquire such a knowledge is via direct experience and contact with actors. CSCW investigated ways to support and foster exchanges by enabling direct communication. This led for instance to the concept of 'Community of Practice' (CoP) [Lave and Wenger, 1991; Wenger, 1998] and 'Community of Interest' (CoI) [Fischer, 2001]. In the first the members share a common practice contributing to a quite homogeneous space whereas in the latter the collaboration is driven by common interests that can bring together diverse heterogeneous backgrounds. Both concepts and their successive derivations address the interactions that enable to capture implicit knowledge of the participants. In a way "*CoIs bring together stakeholders from different CoPs to solve a particular problem of common concern*" [Fischer, 2001]. According to this definition an IPC could be considered as a CoI in a broad sense or better as a collection of dynamically changing CoIs where the common concerns evolve according to scientific goals – "*CoIs often are more temporary than CoPs: they come together in the context of a specific project and dissolve after the project has ended*". For their dynamic and heterogeneous nature the process of learning in CoIs is more challenging than CoPs and requires externalisations and *boundary objects* [Fischer, 2001]. Boundary objects can be seen also as a conjunction point between sharing of knowledge and expertise. For instance, Cabitza et al. propose an approach to promote tacit knowledge by leveraging underspecification in knowledge artifacts [Cabitza et al., 2013, 2008].

In our analysis we adopt such revised concept of boundary object and apply it in a conceptual framework as described in Chapter 5. Additional studies focused on the expertise derived in social contexts such as social networks – 'social capital' [Huysman and Wulf, 2004]. We recognise the importance of those aspects as assets that ought to be preserved and stimulated, thus constituting additional requirements for a collaborative system.

Knowledge and expertise sharing are tightly coupled together, therefore any solution that aims at supporting them must take into account and tackle the associated challenges which have been widely investigated in CSCW. In Section 2.3 we present examples of collaboration platforms that have been inspired by CSCW research.

2.2 Developing agreements

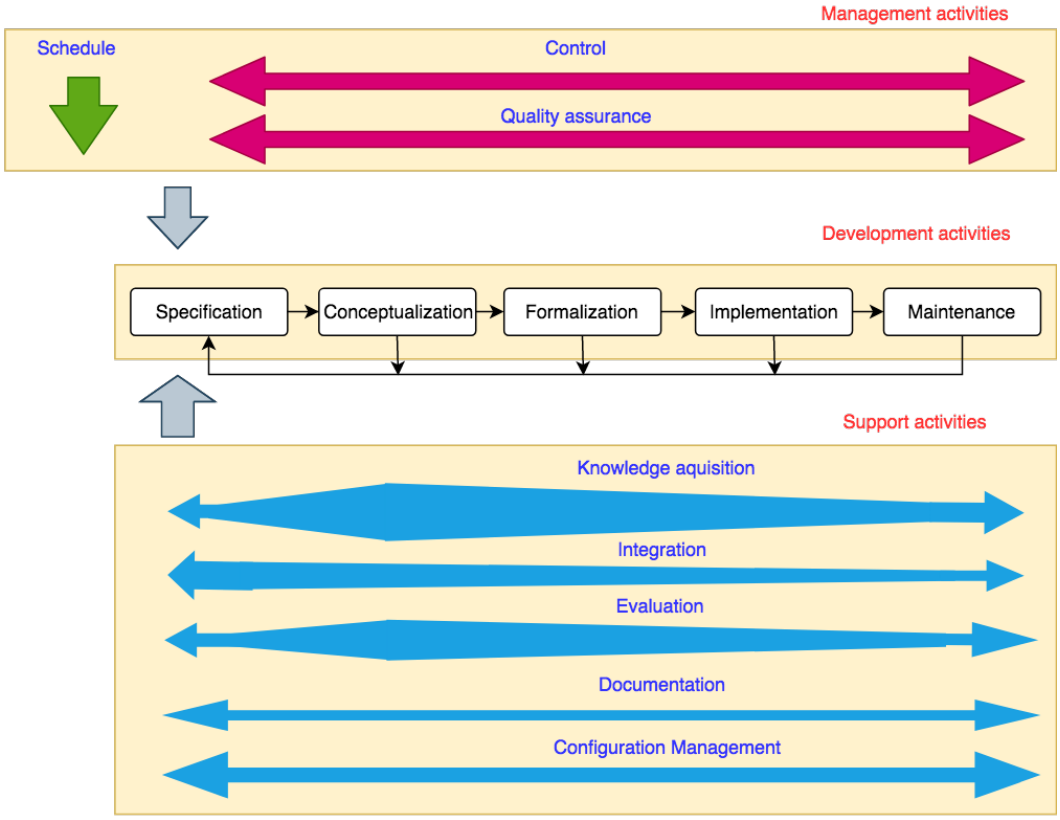
The concept of 'articulation work' applied in the context of CIS captures the processes of establishing shared agreements on the meaning attributed to the collected information. To perform those tasks Bannon and Bødker emphasise the role of 'human mediators' [Bannon and Bødker, 1997]. In Chapter 4 we provide examples and evidence of the socio-technical challenges encountered by those mediators and describe the fundamental role played by organisational frameworks.

Assuming the centrality of the human processes that lead to agreement forming, structured approaches and methods can be leveraged to assist and support them. For instance, an example of such a methodology to achieve semantic agreements is the '*Process and methodology for developing semantic agreements*' by the Interoperability Solution for European Public Administrations (ISA) [PwC EU Services, 2013]. The authors define a process to reach agreement by a 'consensus-building' activity and a methodology to develop and represent semantic agreements. We build on those experiences to develop an approach that targets the dynamics not supported in the methodology by ISA. For instance, we want to understand how we can best sustain the engagement and the active involvement of actors over time. And how we can scale the processes underpinning such agreements when the number of actors and the scale of concerns increase. We will discuss aspects related to the representation of agreements in Chapter 3.

The methodology in [PwC EU Services, 2013] defines three types of stakeholders: an *Authority* that oversees the overall process; *Activity Members* who are in charge of undertaking the plan of action; and a *Wider Community* consulted for feedback and consensus building. It also specifies the roles and the steps of the process that we omit for simplicity. It supports the definition of the overall work-plan, it provides guidelines to help establish a working environment and culture by setting up communication channels, ensuring transparency and record keeping and identifying mechanisms to solve disputes. Particular attention is given to the publication of versions and to the review process with the identification of groups of experts and formal feedback mechanisms. They propose a methodology based on a 'meet-in-the-middle' approach [Zeginis et al., 2014]. That methodology combines a bottom-up approach to collect concepts by "*analyzing the domain and interviewing the domain experts regarding*

their data needs" and a top-down approach whereby existing ontologies are analysed and integrated with the model. For more detail we refer the readers to the ISA document [PwC EU Services, 2013]. We note that standardisation bodies have been using similar approaches to roll out new standards for decades.

The challenges of reaching and formalising agreements have been broadly investigated by the ontology research community – an ontology can be used to represent a common understanding and a community consensus of a domain. In particular in the field of ontology engineering several methodologies have been produced and applied for defining ontologies [Gruninger and Fox, 1995; Fernández-López et al., 1997; Noy and McGuinness, 2001; Corcho et al., 2005; Li et al., 2009; Suárez-Figueroa et al., 2015]. The success of an ontology depends heavily on its adoption within its target communities. Whilst the role of communities is rightly recognised their involvement in the construction phase is not always central. This is typically devolved to ontology experts with domain experts often having some kind of advisory role. Drawing on well-known and popular methodologies such as METHONTOLOGY (Figure 2.1) [Fernández-López et al., 1997; Corcho et al., 2005], Zeginis et al. proposed an approach with "*the active engagement of the domain experts during the actual development of the model (specification and conceptualization) and not just their limited involvement in the model evaluation*" [Zeginis et al., 2014]. They apply such an approach to create the Cancer Chemoprevention Semantic Model (CanCO) and follow 4 phases: *Specification* – "*the scope and the requirements of the semantic model are defined*"; *Conceptualisation* – "*the concepts and relationships of the model are identified*"; *Implementation* – "*the conceptual model is transformed into a computable model using an ontology language*"; and *Evaluation* – it checks "*if the developed semantic model fulfils the requirements defined in the specification phase*" [Zeginis et al., 2014]. In Chapter 5 we introduce an approach that leverages similar principles to build canonical information models underpinning IPC. The examples provided in this section highlight the effort required to achieve agreement about shared definitions. Structured engineering approaches with well-defined steps and phases are essential. However, they often require processes of non-negligible cost, especially in terms of human engagement. The investments for setting up such complex processes ought to be sustained otherwise there is a high risk of disengagement. Indeed, the continued participation of actors is crucial as agreements ought to be maintained and kept up-to-date, tracking the evolu-



Source: [Fernández-López et al., 1997]

Figure 2.1: Example of activities required in a structured process that formalises shared agreements – the METHONTOLOGY ontology engineering approach. It enables the construction of ontologies at the knowledge (*i.e.* conceptual) level

.

tion of the associated information they refer to. Authorities can play an important role by overseeing and steering the activities. However, their establishment is not always possible, *e.g.* due to costs or socio-political reasons. Therefore, it is important to find lightweight and lean approaches that minimise the effort required by the stakeholders, retain their commitments without unnecessary overhead and at the same time guarantee quality and precision of results.

2.3 Platforms for collaboration

Our analysis of the conceptual dimension for information sharing continues by introducing platforms that support and enable collaborative work. We address technical and organisational aspects with a particular focus on *Governance* that is fundamental to promote and sustain collaboration. Examples characterised by diverse typology, context, maturity and target communities are provided. In the next section we introduce Virtual Research Environments (VREs) and similar frameworks *e.g.* Science Gateways (SGs) and Virtual Laboratories (VLs).

2.3.1 Virtual Research Environments and related frameworks

Virtual Research Environments are well-known, powerful frameworks that enable collaborative science. VREs provide scientists and practitioners of communities of practice [Candela et al., 2013] with tools and working environments (or laboratories), usually accessible via the Web, that encompass data, services and computing enabled features such as processing, visualisation, communication, data access and workspaces. Such environments can be deployed in different contexts thereby serving the needs of a variety of communities, however they usually target single disciplines or closely related topics. Recent developments demonstrated the feasibility of aggregating cross-cutting resources to offer VREs as a Service in order to maximise the adoption and productivity in multidisciplinary contexts [Assante et al., 2016a]. Similarly, Virtual Laboratories (VLs), Science Gateways (SGs), Virtual Organisations (VOs) and Digital Libraries (DLs – they are described in more detail in Section 2.4.2) provide the necessary tools and interoperability to enable interactions and foster seamless access, usage and sharing of resources across diverse stakeholders [Gesing and Wilkins-Diehr,

2015; Agosti et al., 2016]. There is a substantial interest in the scientific community in VREs (VLs, SGs, VOs and DLs) that yields a flourishing scientific literature and many initiatives and research projects *e.g.* VRE4EIC[1] and Bluebridge[2]. However, as shown in a recent discussion at the RDA VRE-IG[3], the terminology and the definitions, although often overlapping, are still disputed and often subject to different interpretations. The definition and adoption of reference architectures is an attempt to clarify the focus of each platform and to come to agreed definitions [Pierce et al., 2018; Jeffery et al., 2017].

In our analysis, whilst acknowledging the diverse flavours, we use those terms interchangeably. Such systems deal with the human-computer interactions and socio-organisational issues as well as authorisation and resource management. In this thesis, we assume such a context and focus on supporting the processes needed to build an underpinning alignment of concepts and information.

2.3.2 Virtual Observatories

A model widely applied in geographically distributed independent organisations that share a common research focus is the Virtual Observatory (VO). The concept of Virtual Observatory was first introduced by the astronomers as a means to enable seamless discovery, access and processing of data [National Research Council, 2001; Hanisch et al., 2015]. The goal was to provide an abstraction layer on top of astronomical data provided by independent organisations following the analogy of the World Wide Web. The astronomy community produced a predominant example of successful, long-term collaboration led by the International Virtual Observatory Alliance (IVOA). IVOA discusses and promotes standards for interoperability, protocols for data access and exchange. Since its foundation in 2002 it has supported the astronomy community to establish innovative technical solutions at global scale, disseminate results and promote effective collaborative working practices [Hanisch, 2014].

The IVOA standards roadmap is defined by their Technical Coordination Group (TCG) twice a year. The selection of standards to be promoted is driven by scientific

[1] www.vre4eic.eu

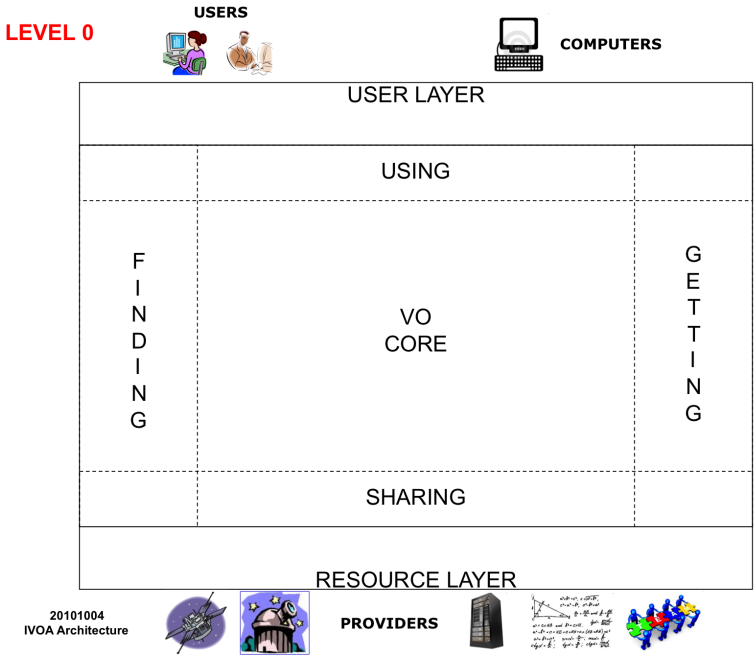
[2] http://www.bluebridge-vres.eu/

[3] https://www.rd-alliance.org/group/virtual-research-environment-ig-vre-ig/post/looking-authoritative-definitions-vre-vlab-science

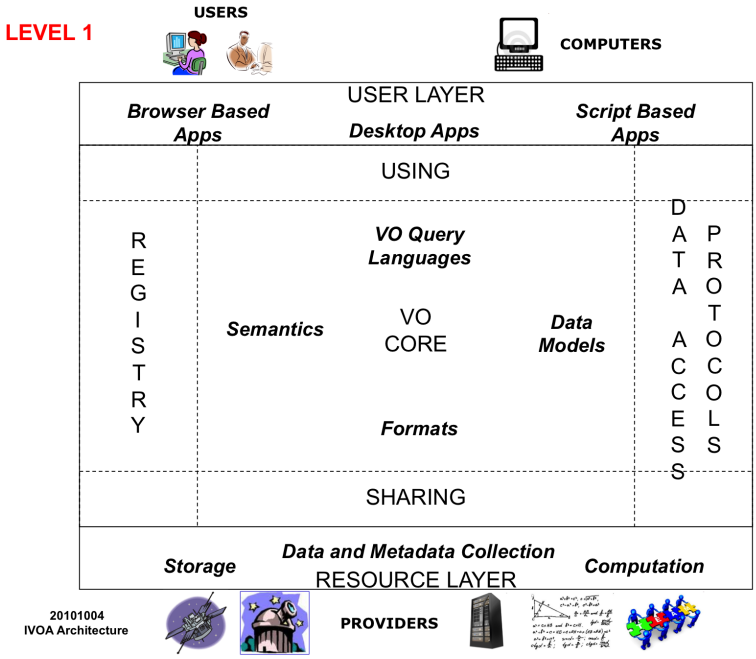
requirements identified by the IVOA Committee on Science Priorities. To address those requirements the IVOA VO offers a technical framework depicted in Figure 2.2. It shows the IVOA architecture which is defined and maintained by the TCG [Arviset and Gaudet, 2010]. The architecture description consists of three levels: Level 0 is a general high-level summary; Level 1 provides details about components and functionalities; and Level 2 couples the supported standards to each component. In Figure 2.2 we present Level 0 and Level 1, we omit Level 2 as the list of standards (nearly 40 in June 2018) therein contained is continuously updated. In Level 0 (Fig. 2.2a) we can appreciate the role of the VO as a technical '*Middle Layer*' that connects transparently the *User Layer* and the *Resource Layer*. The VO enables vertical bidirectional communication in order to find ('Finding') and access ('Getting') resources. Similarly, the role of the standards is to provide horizontal communication between users and providers ('Sharing' and 'Using'). Level 1 (Fig. 2.2b) adds details about features required in the identified high-level functionalities. Each feature is discussed within dedicated groups that address thematic standardisation areas. Currently IVOA investigates and proposes standards for: Data Access, Resources and Registries, Data Modelling and Semantics, Distributed Computational Infrastructure, Collaboration, Authentication and Applications.

The VO model has been successfully applied world-wide and supported by regional and national initiatives. An example is the European Virtual Observatory Euro-VO that "*has been coordinating European VO activities through a series of projects co-funded by the European Commission*" [Genova et al., 2015]. Genova et al. describe the challenges encountered and the coordination activities undertaken in the construction of the Euro-VO. One of their lessons learned is the key role of continued support by programmes and projects that contributed to shape collaborations, to sharpen understanding of goals and approach and to raise awareness in the communities involved and beyond.

In Chapter 4 we present similar findings for the seismology community. These experiences improve our understanding of the powerful role of governance and highlight the significance of a long-term vision to sustain large scientific collaborations. In their analysis of the Euro-VO experience the authors clarified the importance of the support to the scientific community and the data providers. Both have been addressed effectively. For instance, the first has been tackled by setting up a Science Advisory



(a) IVOA Architecture level 0 – It shows the User Layer and Resource Layer which are connected by the services provided by the VO in the Middle Layer



(b) IVOA Architecture level 1 – It shows details about the functionalities of the services enabled in each layer.

Source: [Arviset and Gaudet, 2010]

Figure 2.2: IVOA Architecture

Committee in a very early stage and by fostering direct interactions with researchers *e.g.* with hands-on workshops that yielded valuable inputs for the improvement of standards. The experience acquired in those dissemination activities has been captured as a template – 'Hands-On Workshops' – successively reused at the level of the national VOs. The involvement of the data providers started by performing a census of the available assets and by engaging them in training and dissemination activities. Particular attention was given to raise motivation and to provide incentives to join the VO, for instance, by offering increased visibility and impact [Genova et al., 2015]. Similar strategies or 'rules of engagement' have been harnessed by other communities in different contexts; in [Trani et al., 2018a] we reported our experiences targeting the solid-Earth sciences community. In Chapter 5 we describe those in detail.

The IVOA Virtual Observatory framework has been thought to foster collaboration also beyond the astronomy community. For that scope, the role of standards is crucial. For instance, IVOA resources are made accessible via a Registry of Resources [Demleitner et al., 2014] adopting standard protocols such as OAI-PMH [Open Archives Initiative, 2002] (widely used in the Digital Libraries). IVOA promotes the use of standard vocabularies – "*By adopting a standard and simple format, the IVOA will permit different groups to create and maintain their own specialised vocabularies while letting the rest of the astronomical community access, use, and combine them*" [Derriere et al., 2008]. They suggest the use of RDF and SKOS as standard formats to represent such vocabularies. Those technologies are described in Chapter 3.

Other examples of VOs are: the CLARIN Virtual Language Observatory[4] targeting language resources [van Uytvanck et al., 2012]; and the Web Observatory (WO) – a large system that enables multidisciplinary Web Science [Tiropanis et al., 2013, 2014].

2.3.3 Research Data Alliance

The Research Data Alliance[5] is an international, multidisciplinary, community-driven organisation that is very active in the area of data sharing and exchange, data interoperability and data-driven innovation. RDA focus on both technical and social aspects

[4] `https://vlo.clarin.eu/`

[5] `www.rd-alliance.org`

of data sharing, they envision "*researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society*".

RDA guiding principles are: 1. Openness – meetings, processes and deliverables are public; 2. Consensus – it is achieved among members with proper mechanisms to resolve disputes; 3. Balance – it fosters the participation of balanced representations of members and communities; 4. Harmonisation – it promotes technical and organisational harmonisation; 5. Community-driven – based on a volunteer approach regulated by the RDA Secretariat; and 6. Non-profit – it does not focus on commercial aspects.

Work in RDA is organised and carried out in Working Groups (WGs) and Interest Groups (IGs). The first have a limited time span (typically 18 months) and focus on the delivery of data infrastructures including tools and services. The second have no time limitations and tackle specific issues. IGs can then suggest to set up a WG to develop a solution according to their conclusions. As of February 2018 RDA counts 33 WGs and 58 IGs.

Astronomy was one of the first communities endorsing the approach proposed by RDA and joining the RDA Europe project[6] in order "*to share lessons learnt in the building of the IVOA, and to explore possible liaison with generic interoperability projects*" [Genova et al., 2015].

In RDA recommendations, infrastructure design, policies and various initiatives are emerging to lower the barriers to data, methods and practices sharing and accelerate innovation. Some of these initiatives have recently been endorsed by the European Commission who recognises their importance for referencing in public procurement [European Commission, 2017a], in particular: 1. 'RDA Data Foundation and Terminology Model'; 2. 'RDA PID Information Types API — Persistent Identifier Type Registry'; 3. 'RDA Data Type Registries Model'; and 4. 'RDA Practical Policies recommendations'. The RDA Data Fabric Interest Group introduced the concept of Global Digital Object Cloud (DOC) [Lannom and Wittenburg, 2016] a virtualisation layer that exploits the components presented above to offer an architecture based on the principles of the Digital Object Architecture and fully compliant with the FAIR principles [Wilkinson et al., 2016].

RDA offers us a stimulating environment that influenced this research both at conceptual and technical level. In this section we reported some relevant examples.

[6] `https://cordis.europa.eu/project/rcn/105188_en.html`

2.3.4 Organisations supporting Spatial Data Infrastructures

Prominent examples of organisational models supporting large collaborations have been produced in the context of infrastructures for spatial information. The INSPIRE Directive (2007/2/EC) established a legal framework to share spatial data and support environmental policies in EU [EU Parliament, 2007]. It addresses 34 themes and targets broad and heterogeneous scientific communities. To enable data sharing in those diverse communities has taken "*the best part of 10 years of work to document them through metadata, making the data searchable, viewable and accessible through catalogues and related services*" [Craglia and Nativi, 2018]. A major result of INSPIRE is the harmonisation of policies and rules where achieving interoperability and shared agreements about meanings of concepts has been the biggest challenge. Craglia and Nativi summarise the complexity of that process: "*it was necessary to identify and mobilise the relevant multidisciplinary communities in each of the 34 data themes, and through a patient process of reviewing, refining, and agreeing arrive at shared (generalized) data models that define the structure, content, and meaning of the data needed to support environmental policy. It took some 6–7 years to reach these agreements across hundreds of stakeholder organisations in the member states, and it will take another 10 years to "translate" the existing data in the Member States to the new European models*" [Craglia and Nativi, 2018]. This offers us a clear evidence of complexity, scale and enormous effort required in such endeavours. In the case of INSPIRE an official legal regulation has provided a strong support. However, as we show in Chapter 4, when moving to a global scale it is difficult to achieve such a cohesive and corroborating legal framework.

The Group on Earth Observation (GEO), which is based on a voluntary participation of organisations and governments, coordinates the Global Earth Observation System of Systems (GEOSS[7]). GEOSS is a global initiative to build a large-scale network of content providers into a single overarching system. It embraces the most important existing infrastructures for Earth Observation at a global scale. GEOSS adopts the System of Systems (SoS) approach where several autonomous, independent systems are coherently networked and co-operate to achieve common goals [Jamshidi, 2008]. The GEOSS Platform (former GEOSS Common Infrastructure or GCI) is the

[7] www.earthobservations.org

e-Infrastructure that underpins GEOSS and leverages the distributed independent resources, harmonising data and models, providing access to resources, applications and products. The GEOSS Platform exploits a brokering approach to provide users with transparent access to the distributed resources [Nativi et al., 2015]. GEOSS promotes Data Sharing Principles primarily based on open access to data, resources and services. Currently, the platform manages the access to more than 150 independent data catalogues and information systems.

The concept of SoS captures the common issue of integrating many independent, autonomous systems in order to achieve a global common goal. GEOSS aims to provide decision support tools and what-if type of analysis, with information and knowledge delivery as a goal. Santoro et al. [2016] introduce the Model Web framework that captures business processes as workflows. To address the Science-to-IT barrier issue they leverage models, workflows, vocabularies and knowledge bases. Their focus is primarily on how to combine and use those resources, whereas our focus is on how to support their construction and harmonisation leveraging Core Concepts for collaborations. Our aim is to establish a model to develop consensus and this is not addressed or supported by GEOSS.

2.4 Platforms for data sharing

An important aspect of collaboration is data sharing. Wilkinson et al. summarised nicely the main objectives of scientific data sharing in the formula 'making data FAIR', *i.e.* 'Findable', 'Accessible', 'Interoperable' and 'Reusable' [Wilkinson et al., 2016]. FAIR are not completely new principles; there are several examples of long-established practices addressing similar issues *e.g.* in meteorology, life sciences, astronomy and seismology.

Supported by initiatives such as FORCE11[8], FAIRDOM[9] and organisations such as RDA and the EC FAIR Data Expert Group[10] the FAIR-ness of data has rapidly gained popularity being endorsed in the scientific communities and recognised as a common,

[8] www.force11.org

[9] fair-dom.org

[10] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464

shared goal. In this thesis we assume such a goal, in Chapter 4 we present an approach to enable FAIR principles for the seismology domain. The practical realisation of such principles translates into many different flavours of adherence. However, the challenges faced to establish those principles are recognised and are in common among diverse disciplines [Hodson et al., 2018]. Genova et al. [2017] studied the sharing behaviour in six disciplines including Astronomy and Earth Sciences among others. They identified key common elements for succeeding in that endeavour – "*data sharing should be science driven; defining the disciplinary part of the interdisciplinary standards is mandatory but challenging; sharing of applications should accompany data sharing. Incentives such as journal and funding agency requirements are also similar. For all, social aspects are more challenging than technological ones. Governance is more diverse, often specific to the discipline organization. Being problem-driven is also a key factor of success for building bridges to enable interdisciplinary research*" [Genova et al., 2017].

This process is influenced by different factors and requirements that depend on the context where it is established [Kim and Stanton, 2013; Fecher et al., 2015; Genova et al., 2017]. In the next sections we provide examples of platforms that support data sharing for different purposes. For convenience we present such platforms organising them by their main focus – that is, the principal use case they serve. However, it is important to notice that in practice such divisions might not be so clear and overlaps might be present.

2.4.1 Digital Repositories

Digital Repositories typically support the sharing behaviour for long-term preservation of information. A study over the "European Repository Landscape in 2008" [van der Graaf, 2009] showed an increasing proliferation of research repositories: "*The annual growth rate of the institutional repositories in Europe is 25-35 newly started research repositories per year*". Also, the content hosted in such repositories is very heterogeneous. It focuses mostly on textual forms such as publications or dataset metadata,

but also for instance audio or video. Examples of popular research repositories are: Dryad[11], Zenodo[12], figshare[13], Dataverse[14] and EUDAT B2SHARE[15].

Scientific data repositories are very valuable tools to encourage good data stewardship *e.g.* by promoting well-defined data collection, curation, preservation and dissemination practices [Marcial and Hemminger, 2013]. They often adhere to an open data policy that facilitates the distribution of their contents through integration platforms such as r3data.org[16]. In this way extracting and mining content and identifying relationships can be performed by automated tools. For instance, Manghi et al. propose a toolkit – called D-NET – to generate "aggregative infrastructures" [Manghi et al., 2014]. Aryani et al. set up a system that generates graphs by pulling content from open repositories in order to connect publications, authors, data and grants [Aryani et al., 2018]. They build on the experiences and results of the discussions on integration and interoperability of repositories in RDA. Those are carried out in groups such as the Data Description Registry Interoperability (DDRI) WG[17] and the Research Data Repository Interoperability WG[18]. In Chapter 3 we look at some of the technological outcomes of those WGs as they provide an important support for data sharing and interoperability.

In the area of digital preservation the Open Archival Information System (OAIS) [CCSDS, 2002, 2012] is the blueprint reference model. OAIS was initially developed by the Consultative Committee for Space Data Systems (CCSDS) in 2002 and approved as ISO standard 14721 in 2003. The current revised version has been published in 2012 as ISO 14721:2012. The concept of an Open Archival System is central in the reference model. Despite what the 'open' part might suggest OAIS makes no assumptions about the level of accessibility of its information [Lavoie, 2014]. An OAIS is "*an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make*

[11] datadryad.org

[12] zenodo.org

[13] figshare.com

[14] dataverse.org

[15] b2share.eudat.eu

[16] www.re3data.org

[17] https://www.rd-alliance.org/groups/data-description-registry-interoperability.html

[18] https://www.rd-alliance.org/group/research-data-repository-interoperability-wg/case-statement/research-data-repository

it available for a Designated Community" [CCSDS, 2012]. Therefore, the focus of an OAIS is on preserving information and making it accessible to the broad public but in particular responding to the requirements of communities identified as primary users of the archival system *i.e.* designated communities. Such a key concept captures a user-oriented approach which we endorse and apply in this research.

Figure 2.3 represents the environment surrounding an OAIS. The reference model identifies three main types of stakeholder: 1. Producer – denotes the set of organisations or individuals who provide the information to be preserved in a particular OAIS instance; 2. Consumer – denotes the set of organisations or individuals who interact with the OAIS in order to find and acquire the preserved information of interest. A particular specialisation of this role is the *Designated Community* which is the target set of consumers who should be able to understand the preserved information; and 3. Management – "*The role played by those who set overall OAIS policy as one component in a broader policy domain, for example as part of a larger organization*" [CCSDS, 2012]. It is not involved in the daily administration which is delegated to another internal component (*Administration*).

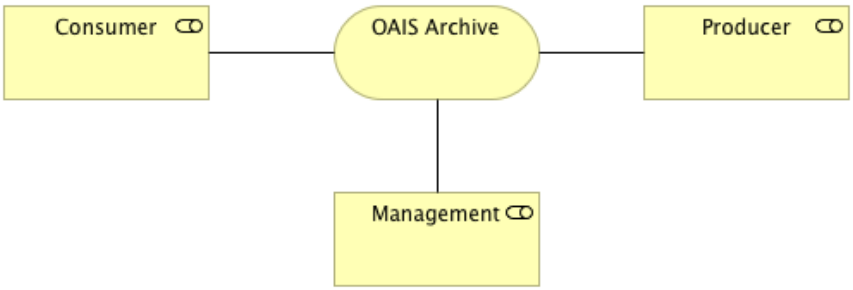


Figure 2.3: Representing the environment surrounding an OAIS. Three main types of roles interact with the archival system: Consumer, Producer and Management.

The reference model identifies a minimum set of mandatory responsibilities that characterise an OAIS. Those are described in [CCSDS, 2012] and listed below:

- "*Negotiate for and accept appropriate information from information Producers*".
- "*Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation*".

- "*Determine [...] which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base*".
- "*Ensure that the information to be preserved is Independently Understandable to the Designated Community*".
- "*Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive*".
- "*Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity*".

The OAIS RM is a high-level model which provides a quite detailed description of functionalities, components and roles of such an information system. The model does not prescribe implementation details or technical specifications. The lack of concrete specifications makes it difficult to formally validate compliance with the RM. Compliance is often achieved with different gradations and assessed for instance by mapping functionalities of real systems onto the ones of the RM [Vardigan and Whiteman, 2007] – "*Conformance to the reference model can imply an explicit application of OAIS concepts, terminology, and the functional and information models in the course of developing a digital repository's system architecture and data model; but it can also mean that the OAIS concepts and models are recoverable from the implementation*" [Lavoie, 2014]. The freedom to conveniently derive elements form the model is also a reason of its success, widespread adoption and application in several systems [Hou et al., 2014; McDonough, 2011; Brunsmann et al., 2012; McMeekin, 2011].

In the context of this research the OAIS RM has been an important source of inspiration, we adopt and build on several of its features and concepts. The model indicates that information must be preserved in such a way to guarantee continued understanding and usage by OAIS' designated communities. This high-level target requirement has direct implications in the OAIS information model that describes the types of elements necessary to enable those functionalities. We build on such

information model and its application in catalogues. In Chapter 3 we will provide more details about representation aspects. Also, we leverage concepts such as the interaction models of distributed OAIS, interoperability of archives and governance.

The OAIS RM assumes that information is packaged and preserved in units that are frozen at the moment of submission to the system. In our case we target an open information space that evolves dynamically to incorporate changes required by the designated communities' working practices. We focus on those dynamic processes that are out of scope in the OAIS context. For this reason, not only do we recognise the centrality of the user communities but we also envision a stronger participation in the selection and maintenance of the content to be preserved. They need not only to influence the system with their requirements but be actively engaged throughout the lifetime of the system – they ought to feel responsible for the content managed in the shared space. In the next section we move towards systems that in the recent years have gone through an extraordinary evolution and reached high levels of complexity: Digital Libraries.

2.4.2 Digital Libraries

Digital Libraries' (DLs) main focus has been for a long time the collection, organisation and publication of digital content, thus supporting the sharing behaviour for publication. Although not as their primary focus such a behaviour can be exposed also by Digital Repositories presented in the previous section. Assante et al. analysed aspects of data publication in generalist scientific repositories and assessed their limited support for data management and usage of their resources [Assante et al., 2016b]. They studied functionalities considered as fundamental for data publishing: dataset formatting, documentation, licensing, publication costs, validation, availability, discovery, access, and citation. DLs tackle those issues and provide effective technical solutions to enable such features. They are complex information systems that extend their scope beyond preservation and can play a significant role in data management, publication and sharing practices. DL are often multidisciplinary and heterogeneous, thus handling many types of digital objects, but in some cases they can be tailored to meet specific communities' requirements. DLs have been deployed and adopted in several contexts and domains such as healthcare [Kostkova and Madle, 2013], education and scholarly com-

munication and scientific research. Borgman et al. performed an interesting analysis of four research sites spanning from 'big data' sciences, such as astronomy, to 'little' sciences, with disciplines such as engineering, life sciences and physical sciences, in order to understand the role of DLs as scientific knowledge infrastructures [Borgman et al., 2015]. Of course, today such systems can involve massive volumes of data; the sensitivity or commercial-in-confidence nature of their data is a greater differentiator. That study shows the potential of DLs in support of the different phases of the data management from data collection to curation, preservation and publication. DLs encourage and foster data sharing practices offering technological platforms and support especially in those contexts where *ad hoc* solutions are not affordable.

In recent years DLs have shifted from 'content-centric' to 'person-centric' systems thus targeting users' experience and facilitating communication and collaboration [Candela et al., 2011]. This "*vision of Digital Libraries seems to resonate well with the concept of 'Information Space'*". It suggests that we investigate those systems as a valuable source of inspiration for our research. In this new light DLs enable the concept of 'Inhabited Information Space' where "*both information and people who are using that information (viewing it, manipulating it) are represented*" [Candela et al., 2011]. Candela et al. in their Digital Library Manifesto recognise the complementary role of DLs and CSCW research – "*Digital Library provides an Information Space that is populated by a user community and becomes an Inhabited Information Space through CSCW technology*".

Therefore, DLs have evolved into systems heterogeneous in their targets and scopes that offer a wide rage of functionalities. For this reason it is difficult to find a comprehensive definition to characterise them. Attempts to clarify the role of DLs and their definition yielded conceptual frameworks such as the 5S Framework [Gonçcalves, 2004] and the DELOS Reference Model [Candela et al., 2007]. Agosti et al. proposed also a model to make them interoperate [Agosti et al., 2016]. Further refinements led to the the DL Manifesto that describes a DL framework constituted of three components as illustrated in Figure 2.4.

A Digital Library is defined as a "*potentially virtual organisation, that comprehensively collects, manages and preserves for the long depth of time rich digital content, and offers to its targeted user communities specialised functionality on that content, of defined quality and according to comprehensive codified policies*". A Digital Library

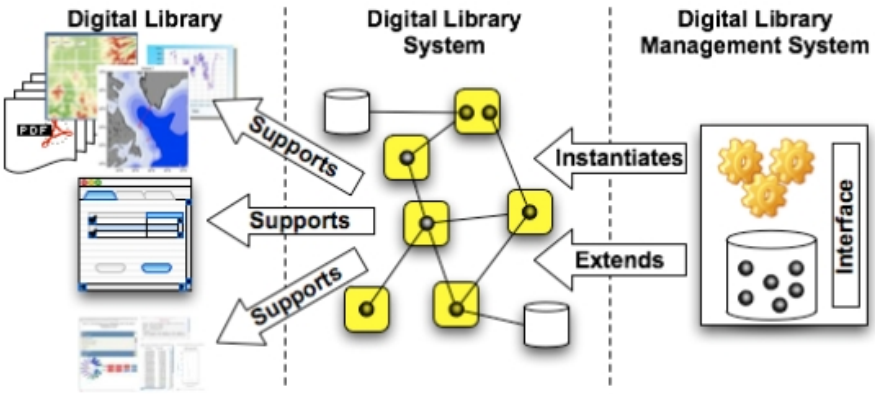


Figure 2.4: A three tier architecture of a Digital Library framework

Source: [Candela et al., 2011]

System (DLS) "*software system that is based on a possibly distributed architecture and provides all facilities required by a particular Digital Library*". It enables the interaction with the users. A Digital Library Management System (DLMS) "*software system that provides the appropriate software infrastructure both (i) to produce and administer a Digital Library System incorporating the suite of facilities considered fundamental for Digital Libraries and (ii) to integrate additional software offering more refined, specialised or advanced facilities*".

The model identifies seven core concepts as foundations for a DL.

1. Organisation that surrounds the DL – "*it is a social arrangement pursuing a well defined goal*".

2. Content – it "*encompasses the data and information that the Digital Library handles and makes available to its users*".

3. User – individuals and groups, "*the various actors (whether human or machine) entitled to interact with Digital Libraries*".

4. Functionality – it encompasses the services offered by the DL (*e.g.* registration of new information objects, search, browse).

5. Policy – it "*represents the set or sets of conditions, rules, terms and regulations governing every single aspect of the Digital Library*" *e.g.* concerning usage of resources, digital rights and privacy.

6. Quality – it "*represents the parameters that can be used to characterise and evaluate the overall service of a Digital Library*".

7. Architecture – it targets the DLMS and "*represents a mapping of the overall service offered by a Digital Library (and characterised by Content, User, Functionality, Policy and Quality) onto hardware and software components*".

We notice a certain overlap with the OAIS RM that certainly inspired research in DLs and provided solid conceptual foundations. Those intellectual investments and efforts are valuable influences for our research as they target similar conceptual spaces. For instance, the concept of 'content' in a DL comes with associated policies and agreements. They usually refer to the negotiation process with the organisation underpinning the DL. DLs typically do not aim to influence the content of information provided but rather to offer best possible ways to collect and to use it. Therefore, promoting agreements and a shared vision among their users is not their primary goal. Nevertheless, such a collaborative behaviour might emerge as a consequence of the sharing behaviour. In this research we target shared agreements as a primary goal. This requires an active engagement strategy with the participants that has to be sustained. We recognise the value of the conceptual and technological advances of DL research and the strategies successful in specific contexts such as scholarly communication. We build on those to pursue our research goals. In the next section we introduce platforms that promote data sharing by offering a common abstraction that facilitate processing and analysis: Data Cubes.

2.4.3 Data Cubes

The platforms introduced so far focus predominantly on managing and providing access to data – they support their users, for instance, by offering finding aids. Often the methods that enable an effective usage of data are not in their primary concern – they remain in the realm of the users.

In this section we complement that picture with a slightly different view that brings in elements of data usability, methods and related requirements within an integrated system, thereby enabling the sharing behaviour for data processing and analysis. Several platforms support such a behaviour, for instance, some DLs can

be harnessed as 'intelligent' systems that facilitate the generation and enactment of automated workflows [Leidig and Fox, 2014].

Here we focus on a particular category of such platforms which draws on the abstraction of '*data cube*'. This is interesting in the context of this thesis because it provides an example where a conceptual view of data, *i.e.* a data cube, can be applied successfully in real systems to achieve agreements, *e.g.* about data structure, access services and operations. Also, in some domains data cubes foster cross-disciplinary collaborations by promoting the use of standards [Nativi et al., 2017].

The concept of data cube appeared in the 1990s in the Business Intelligence domain. In that context it was typically applied in Data Warehouse systems to represent a two or three-dimensional dataset *e.g.* a table, a spreadsheet. The motivating idea was to facilitate analysis by organising the target data according to a defined and easy-to-handle structure. The data cube abstraction and the associated operations were embedded and supported natively in databases known as Online Analytical Processing (OLAP) [Codd et al., 1993; Gray et al., 1997].

More recently the Datacube Manifesto defined a data cube as: "*a massive multi-dimensional array, also called 'raster data' or 'gridded data'; 'massive' entails that we talk about sizes significantly beyond the main memory resources of the server hardware. Data values, all of the same data type, sit at grid points as defined by the d axes of the d-dimensional datacube. Coordinates along these axes allow addressing data values unambiguously*" [Baumann, 2017]. Building on that definition Strobl et al. identify six main dimensions that characterise data cubes [Strobl et al., 2017]. They refer to Geospatial Data Cube (GDC) to describe a system "*based on regularly and irregularly gridded, spatial and/or temporal data with n dimensions (or axes) and characterized by the presence of the 6 faces*". Such faces correspond to the features that a GDC enables and are:

1. Parameter Model – describing the semantics of the cube cell.

2. Data Representation – describing how a parameter is discretised and encoded along the axes of the cube.

3. Data Organisation – dealing with the physical arrangement of the discretised parameters.

4. Infrastructure – hosting the data storage units.

5. Access and Analysis – providing functionalities to manipulate the cube via APIs.

6. Interoperability – enabling the fusion of different spatial information.

The latter depends on the broad adoption of standards that can be fostered by data cube infrastructures. Nativi et al. focus on interoperability aspects and propose a view-based model on top of data cubes [Nativi et al., 2017].

Data cubes are gaining popularity to address the challenges of geospatial computations that often involve large amounts of data. For instance, when the subject of investigation are regions with large extensions and the desired analysis targets high resolutions, the volume of input/output data to handle increases ineluctably. Geospatial data are characterised by projections and coordinate reference systems in which the spatial components are encoded. Combining and sub-setting such datasets (even with the same type of observations) might not be straightforward. Those operations demand specific knowledge of underlying details such as reference systems and encodings. As a consequence the management and data manipulation processes can be time consuming and error prone. Recent developments propose a data cube approach to address those issues. Data cubes offer several improvements and foster the use of standards to interact with geospatial data, *e.g.* the guidelines of the Open Geospatial Consortium (OGC) [Maidment et al., 2011]. Data cubes are also associated with the concept of Analysis Ready Data (ARD) and have been adopted to perform the analysis of large time series (*e.g.* satellite observations) and to enable real time exploration and visualisations [Lins et al., 2013]. Recently, major initiatives exploiting data cubes have been launched to address the big data challenges of different scientific communities. EarthServer[19] enables Big Earth Data analytics on a variety of integrated products [Baumann et al., 2016, 2018]. The Open Data Cube Initiative[20] promotes an open and collaborative data-cube approach to maximise the value and impact of satellite observations. Earth System Data Cube (ESDC[21]) by ESA focuses on the detection of effects of climate change in terrestrial ecosystems. Finally, the Joint Research Centre

[19] `http://www.earthserver.eu/`

[20] `www.opendatacube.org`

[21] `http://earthsystemdatacube.net`

Earth Observation Data and Processing Platform (JEODPP[22]) supported by the European Commission combines high-performance computing and petabytes of scalable storage to analyse satellite and earth observations. A noteworthy initiative is the W3C Data Cube Vocabulary that offers a RDF representation widely used for statistical data [Cyganiak and Reynolds, 2014].

The data cube is a powerful abstraction that offers an approach to organise information in a common structure along homogeneous dimensions. Concepts with a different semantics might share a common structure, for instance, a data cube might represent data containing Earth observations [Baumann et al., 2018] or data for Business Intelligence [Codd et al., 1993].

Despite the differences in implementation and semantics by sharing a common structure data cubes enable shared operations such as subsetting, projecting *etc.* Of course, the semantics of those operations might be different depending on the concept represented in the cube. Such platforms promote collaboration that derives from sharing a common infrastructure with shared functionalities for data management, operations and services. Although the support for interaction and mutual exchange of concepts among different groups of users might be limited and not necessarily their primary focus; the conceptual partitioning of those platforms along different dimensions [Strobl et al., 2017] enables a separation of concerns. Users can focus on specific aspects such as 'Infrastructure' and join efforts to find common solutions. For instance, by separating the concepts of *coverage*, adopted to model data, and the *service* model any compliant standard interface can be used to consume the data in a data cube [Baumann et al., 2018]. Building on this conceptual view standards such as OGC Web Coverage Service (WCS) and Web Coverage Processing Service (WCPS) have emerged and have been broadly adopted *e.g.* in the European legal framework for Spatial Data Infrastructures, INSPIRE [INSPIRE Maintenance and Implementation Group (MIG), 2016]. In particular, WCPS embeds elements of computation in the query mechanism allowing systems to delegate (part of) the processing to the data cube platform, thus local to the data. This aspect is very important as the costs of data movements are becoming increasingly unaffordable.

Therefore, data cubes provide valuable contributions to building the conceptual space of this thesis that can be summarised as follows: a) a powerful conceptual

[22] https://cidsecure.jrc.ec.europa.eu/home/

metaphor; b) partitioning of the concerns into independent dimensions; and c) identification of standards as key collaboration-enablers.

We recognise some limitations in the data-cube approach (*e.g.* the need to modify data structures to align with the cube and consequently the need to adjust existing methods), nevertheless we leverage results and lessons learned to elaborate our strategy illustrated in Chapter 5.

2.5 Summary and conclusions

The goal of this chapter is to lay foundations of the conceptual space underpinning our research. We pursued that goal by reviewing distinct bodies of literature and by reporting and analysing aspects and results that are relevant for this thesis. Our strategy has been to move progressively from high-level broad views to more focused examples of applications. We started with the perspectives about collaborations and knowledge management provided by the rich CSCW literature. That context offers us key concepts and a powerful terminology. It provides us a deep analysis of socio-technical aspects and a broad range of conceptual tools. For instance, we learned from the CSCW about effective ways to represent and share knowledge in a collaborative environment. We realised that applying a repository model is in practice less beneficial than creating and maintaining a Common Information Space [Bannon and Bødker, 1997]. A CIS can represent a wide range of concepts that are relevant for the participating parties who can contribute and exchange their knowledge by means of 'boundary objects' [Star and Griesemer, 1989] – those powerful abstractions can be also used to capture expertise and tacit knowledge [Cabitza et al., 2013].

We recognised the importance of shared agreements to sustain the conceptual definition of a CIS. Leveraging the notion of 'articulation work' and the central role of 'human mediators' [Bannon and Bødker, 1997] we moved towards the analysis of best practices and methods to establishing agreements. The ontology engineering domain and the standardisation bodies provide us effective methodologies to build and formalise such agreements [Fernández-López et al., 1997; PwC EU Services, 2013; Zeginis et al., 2014]. We realised the value of an active engagement of the target actors over time – that is often not sufficiently sustained after the definition of the shared agreements.

We examined diverse platforms that enable collaboration with a particular focus on their organisational structures and governance models and recognised the value of VREs, SGs, VLs, *etc.* We appreciated the concept of Virtual Observatory, the achievements and the advances of organisations such as the IVOA [Genova et al., 2015]. They show how the astronomy community successfully tackled complex challenges by pioneering and establishing effective methods for global collaboration. Astronomers contributed their lessons learned and experiences as initial supporters of the RDA. We recognised the great value of such a community-driven initiative and reported approaches applied by the spatial data infrastructures communities to develop and achieve agreements in INSPIRE and GEOSS. Those initiatives provide a clear picture about scale and complexity of the challenges addressed by this research.

After reviewing conceptual and organisational models supporting collaborations, we focused on a critical aspect underpinning them: data sharing. We analysed three types of platforms enabling data sharing: Digital Repositories, Digital Libraries and Data Cubes. They all contribute interesting and relevant perspectives by addressing sometimes complementary needs. We reported about the increasingly popular FAIR principles that provide solid conceptual foundations and usable tools for data sharing. We acknowledge their great value but at the same time we recognise that there are still a number of open issues and barriers to overcome for their effective application. A culture is needed that establishes those principles effectively [Hodson et al., 2018]. Hodson et al. also recognise the need for 'disciplinary interoperability frameworks' to support science-driven developments of the data-sharing behaviours – in this way communities are motivated and incentivised [Hodson et al., 2018]. To empower collaboration in large federated environments interactions leading to agreed definitions and expertise sharing should be equally sustained.

To conclude our analysis of the scientific literature reviewed in this chapter we present in Table 2.1 a summary of: a) contributions provided to our research goals; and b) open issues and identified gaps.
The first provide elements on which we build our strategy whereas the latter motivate our research goals.

Table 2.1: Summary of literature contributions

Element	Lessons learned	Open issues
CSCW	CIS, Conceptual articulation and tools	Focused application scope, mostly targeting humans
Agreements development	Ontology engineering, Methodologies	Limited user engagement and processes to sustain it
VREs, SGs, VLs ...	Established heterogeneous collaborative environments	Pre-shaped, limited user's influence on requirements, mainly targeting tools
VOs	Governance, shared practices, successful model	Typically targeting single disciplines; require tight working relationships
OAIS	Conceptual model	Focused on preservation of closed units of information
FAIR	Endorsed approach, generic	It addresses only the technical part of the challenges and requires support for boundary research and innovation
DL	Established frameworks, moving towards CIS	Focused scope
Data cube	Established practices, foster collaboration by separating concerns	Suitable for specific data types and use cases; limited interoperability

Later in this thesis some of the elements presented in Table 2.1 will reappear applied in different contexts. After investigating the conceptual dimension in the next chapter we move towards other two aspects – we look at how to represent concepts and how to build and maintain populations of concepts.

Chapter 3

Representations and Populations for Common Information Spaces

In this chapter we continue our literature review by addressing the remaining two dimensions introduced in Chapter 1, namely Representation and Population. In the first part we review approaches and methods to represent information and knowledge, we then proceed by reporting about tools, protocols and frameworks to instantiate, manage and exchange populations of concepts. Finally, we present a selection of examples of integrated systems designed to fulfil relevant application scenarios targeting heterogeneous, distributed data sources.

We build on the conceptual view introduced in Chapter 2. For instance, reprising the concept of Common Information Spaces, here we focus on the effective construction of such spaces.

3.1 Representing Information

In the previous chapter we looked at how to define conceptual spaces underpinning collaborations, and we appreciated the complexity and the effort required to form and maintain agreements. Now we start digging into more technical aspects by investigating ways to represent such spaces, their implications and challenges.

We review methods to model concepts and their relationships as well as notations to express them targeting both humans and automated systems. Our interest is mainly on the latter in order to support the establishment of automated methods.

Information can be represented in many ways and multiple representations might exist for the same piece of information. For instance, the two sentences: "*My name is Luca Trani*" and "*Il mio nome è Luca Trani*" represent identical concepts and convey the same information in two distinct languages: English and Italian respectively. Two sequences of characters carry equivalent meaning. Likewise, we might represent similar concepts in a graphical notation as in Figure 3.1:

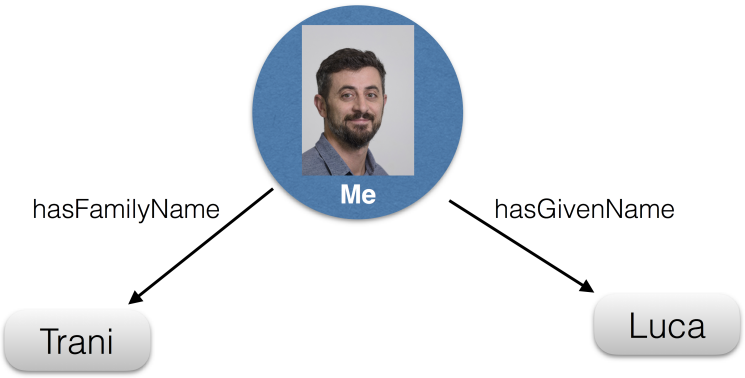


Figure 3.1: Expressing information with a graphical representation.

These examples illustrate representations that are primarily suitable for human interpretation. To support automated systems additional characteristics and features might be needed, they are discussed in this chapter. For instance, graphs are powerful models to represent information and knowledge. They can be visualised as a set of nodes or vertices connected by edges – the example illustrated in Fig. 3.1 is a graph. Also, as they are mathematical models, they can be expressed with precise notations, *e.g.* $G = (V, E)$, and their behaviours can be described by formalised theories. This allows us to effectively adopt graphs both for human communication and machine interpretation. In this chapter we come back to these representations and introduce specific types of graphs adopted in the Semantic Web [Berners-Lee et al., 2001]. In the next section we leverage the OAIS RM to draw characteristics of representations by observing their application in digital preservation.

3.1.1 The importance of description in digital information

The notion of representation assumes particular importance when associated with data and digital information. It is a means to make characteristics explicit, expose be-

haviours and consequently enable actions on data. Hence, data and representation are tightly bound together. For instance, the FAIR principles target data but actually they provide requirements for their representation – data should be described, *i.e.* represented, in such a way that they can be discovered, accessed, interpreted and used. Each of those functionalities poses specific requirements on the corresponding representation, *e.g.* they might yield different approaches when targeting humans or automated tools. FAIR principles are meant to be broadly applied, they do not impose constraints about target users.

The OAIS RM encompasses additional aspects that influence the representation associated with data. In OAIS a major goal is to ensure that data preserved in an archive should be interpretable and usable by the identified *designated communities* in the *long-term*. Such a challenging goal brings in two key viewpoints:

1. identification of target users – restricts the problem space by providing a clear focus but implies the characterisation of those user communities; and

2. identification of an indefinitely large temporal horizon – requires archives to cope with undefined and unpredictable scenarios, *e.g.* technologies and approaches might radically change but the preserved information must stay valid and consistently usable.

To address those challenges the OAIS RM defines an information model that includes the components required to accompany any information object deposited in an OAIS archive. In the next section we introduce elements extracted from the OAIS information model that will help us understand and address the representation dimension.

3.1.2 OAIS Information Model

Some preliminary definitions from the OAIS RM [CCSDS, 2012] are required to better understand its information model. We present them below.

Definition 2. *"A person, or system, can be said to have a **Knowledge Base**, which allows that person or system to understand received information. For example, a person who has a Knowledge Base that includes an understanding of English will be able to read, and understand, an English text".*

Definition 3. ***Information*** *is defined as "any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data in an exchange".*

Definition 4. ***Information Object*** *is composed of a* ***Data Object*** *and* ***Representation Information*** *"that allows for the full interpretation of the data into meaningful information".*

A Data Object can be a Physical Object (*e.g.* a rock sample) or Digital Object (*i.e.* a sequence of bits). The Representation Information accompanying a Data Object provides additional meaning. For instance, in the case of a Digital Object, it maps the bits into commonly recognised data types such as character, integer, real and into structures of these data types. It can also include the description of interrelationships between objects. In the case of a Physical Object the Representation Information includes the known characteristics of the object derived for example from an analysis.

To fulfil its goals "*the OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained*". This requires important governance decisions about "*maintaining the minimum Representation Information needed for its Designated Community, or maintaining a larger amount of Representation Information that may allow understanding by a larger Consumer community with a less specialized Knowledge Base*" [CCSDS, 2012]. Analogous governance implications are reprised in Chapter 5.

The Representation Information is composed of diverse elements. **Structure Information** describes format and structure of data *e.g.* "*common computer data types, aggregations of these data types, and mapping rules which map from the underlying data types to the higher level concepts*". The Structure Information is often referred to as the *format* of the digital object. "*The Representation Information provided by the Structure Information is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Information, the additional information as to which language was being expressed should be provided. This type of additional required information is referred to as the* ***Semantic Information***". In scientific data "*the information in the Semantic Information can be quite varied and complex. It will include special meanings*

associated with all the elements of the Structural Information, operations that may be performed on each data type, and their interrelationships" [CCSDS, 2012].

The Semantic Information is independent of the format; for instance the meaning of some words in a text is independent of whether it is encoded in Word or PDF. Figure 3.2 depicts elements of the OAIS information model. We can notice the presence of an additional element: **Other Representation Information**. This component captures the missing concepts that cannot be directly related to Structure or Semantic Information. For example, information on how to relate Structure and Semantic, processing or algorithms, software or any other information which may be needed to interpret the Data Object.

Representation Information may be itself composed of other Data Objects with their associated Representation Information – the resulting set of objects form a **Representation Network**.

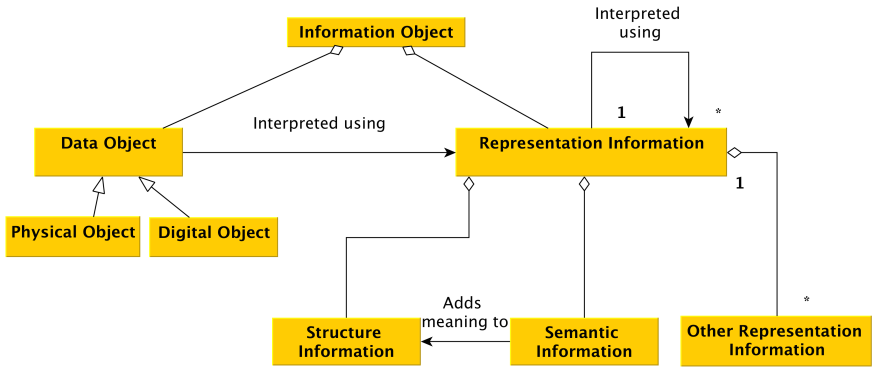


Figure 3.2: Class diagram illustrating key concepts of the OAIS RM Information model and their relationships: Information Object and Representation Information

Source: [CCSDS, 2012]

The concept of Representation Information is crucial to preserve the intended meaning of a Data Object and to enable its interpretation. Being an Information Object in itself it can be associated with digital or physical forms. In the former case the recursion of Representation Information, contained in a Representation Network, eventually should lead to physical forms which can be understood by a designated community. The adoption of known forms, such as textual descriptions in well-known standards like UTF-8, can facilitate the preservation of Representation Information. Standards

and formal description languages defining the constructs and data structures can help resolving ambiguities. Such languages may require additional textual description to convey the intended meanings of the Representation Information. In the next section (3.1.3) we explore examples of formal descriptions, *i.e.* metadata.

OAIS characterised Information Objects based on their content and role played in the long-term preservation context – they defined a taxonomy which is illustrated as an example in Fig. 3.3. Similar classifications have been defined with metadata as we show in the next section.

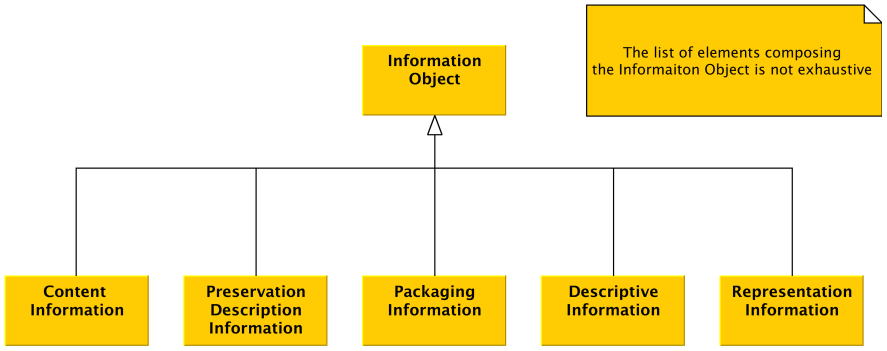


Figure 3.3: Categories of Information Objects defined by content and function in OAIS operations.

Source: [CCSDS, 2012]

The OAIS RM identifies another important point: the dynamic nature of the Knowledge Base for a designated community. The Knowledge Base varies over time thus requiring a periodic update of the Representation Network. The OAIS RM envisages two possible ways to maintain the Representation Network of an OAIS archive and leaves the final choice to the implementation phase: *collecting* all the Representation Information or *referencing* to trusted or partner OAIS archives. We will come back to these aspects when we address the population dimension later in this chapter.

In conclusion, we showed how the OAIS RM provides us with a rich set of conceptual tools that can help us model the representation aspects. OAIS focus is primarily on preservation but its concepts can be generalised extending their scope. For instance, the Information Object with its Representation Information provides a powerful logical model that decouples a data object and its representation – in this

thesis we leverage those concepts. In the next section we tackle the issue of how to build formal descriptions, *i.e.* representations, with metadata.

3.1.3 Metadata

The OAIS RM defines an Information Object and identifies the components required for interpreting and using it. In this section we introduce the concept of metadata that can be adopted as a means to construct such components thereby achieving information representations.

Metadata are typically known as '*data about data*'. This is a quite generic and broad definition that arguably provides useful information for an effective application and exploitation of metadata. A fairly established approach suggests to define metadata according to the functions they enable. For instance, the US National Information Standards Organisation (NISO) defines metadata as "*structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource*" [NISO, 2004]. Therefore they identify functions such as: description, identification, discovery, retrieval and management. Also, the application of metadata can be perceived differently depending on the context – whilst in modern systems metadata are predominantly associated with machine readable information, there are cases where they might be intended for human interpretation (*e.g.* in a manually annotated medical record). Metadata are at the core of many information systems and serve a wide variety of purposes. They can be categorised based on their applications. For instance, a typical classification originated from the cultural heritage community is described in [Riley, 2017] and defines the following types of metadata:

- **Descriptive** – "*information about the content of a resource that aids in finding or understanding it*". It contains information such as identifier, author, publisher and subject.

- **Administrative** – "*information needed to manage a resource or that relates to its creation*". It includes:

  - **Technical** metadata – "*information about digital files necessary to decode and render them, such as file type*".

  - **Preservation** metadata –"*supporting the long-term management and future migration or emulation of digital files, for example, a checksum or hash*".
  - **Rights** metadata – "*such as a Creative Commons license, which details the intellectual property rights attached to the content*".

- **Structural** – "*describe the relationships of parts of resources to one another*". For example, in a book it provides information about pages, chapters, sections, table of contents *etc.*

- **Markup languages** – "*Integrate metadata and flags for other structural or semantic features within content*". Examples are markups in a textual resource and flags that highlight notable content. As we discuss later such languages can be harnessed to enable collaborative participation in the metadata creation process *e.g.* via users' annotations.

Another type of classification identifies *static* and *dynamic* metadata, thus focussing on the variability of the information they represent. Such classifications might help us understand the use of metadata but those categories might not be exhaustive and introduce limitations – some purposes (*e.g.* long-term preservation) require metadata that span multiple categories.

In the context of this thesis, unless otherwise specified, we use the term metadata in a broad sense to indicate a formal description or representation accompanying an Information Object.

Metadata are often organised in a '*metadata schema*' which contains sets of concepts or *metadata elements*. Such schemes might yield *metadata standards* developed to address the requirements of a particular domain or concern, as a consequence the two terms are often used interchangeably. In Section 3.3 we introduce some relevant metadata standards as they provide examples of well-established representations. Often they result from broad agreements and continued collaborative efforts. Therefore, as we discuss in Chapter 5, it is important to retain the value of those intellectual investments and apply re-use as a principle whenever possible.

Metadata can be stored, maintained and exchanged in a variety of forms and *encodings*. Popular tools to store metadata are files, databases and catalogues whereas they are usually exchanged in textual forms encoded in languages such as XML or JSON. In the next sections we provide examples of such encodings.

The value of metadata is widely recognised in the management of scientific data throughout their lifecycle [Gray et al., 2005]. They support automated workflows, computation and visualisation, they can be used to record provenance and enable reproducibility. Also, they enable *interoperability*: "*the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality*" [NISO, 2004] – this is discussed in Section 3.3.5.

Later we look into more details of some of those aspects as they are relevant to our analysis.

3.1.4 Structuring metadata

We have seen how metadata can be exploited for a wide range of purposes and enable different functionalities. Metadata serving a common purpose are typically organised and documented in a consistent structure thereby defining a *metadata schema*. A schema contains a set of terms or *metadata elements* with their names, definitions and meanings. The elements of a metadata schema constitute its *vocabulary*. The value associated with a metadata element is known as *content*. A metadata schema might also include content rules – specifying how the content should be provided, *e.g.* allowed values – and syntax rules – specifying the type of encodings or formats. Several organisations oversee the standardisation and maintenance processes of metadata schemes. Examples are ISO[1], W3C[2] and OGC[3]. Also, community-based initiatives promote the use of domain or application specific standards – in Chapter 4 we provide an example of such a community effort in the Seismology domain.

Metadata elements from one or more schemes can be combined to fulfil the requirements of a specific application scenario. The application of metadata elements for a specific use can be expressed with an *Application Profile* [Heery and Patel, 2000] – it contains (sub-)set of elements of metadata schemes with guidelines and rules to express which values are valid for the intended context. Similar to metadata standards, application profiles can be documented and formalised. Later in this chapter we provide examples of notations that allow us to define and exchange application profiles.

[1] www.iso.org/

[2] www.w3c.org/

[3] www.opengeospatial.org/

Metadata can be structured and encoded according to different models *e.g.* relational, hierarchical/tree-based or graph-based. Depending on the type of information they represent some structures might be more suitable than others. For instance, when the focus is on describing and capturing relationships among elements and their dynamics graph-based structures might be indicated. Tree-based (a particular type of graph) forms might fit the purpose of describing hierarchies whereas tables can be used for static information such as size and path of a file.

Relational representations are popular in databases (addressed in Section 3.4.1), in that context a collection of metadata is called *record*.

XML metadata standards are examples of tree-based models. A wide variety of standards offer XML encodings – their elements are typically defined using the XML Schema Definition Language (XSD) [Gao et al., 2012]. XSD specifies entities, attributes, constraints and enables validation.

RDF is a well-known graph-based representation which recurs often in this thesis and it is described in the next section.

3.1.5 Semantic Web and Linked Data

The Resource Description Framework (RDF) is one of the technological pillars of the Semantic Web [Berners-Lee et al., 2001]. The revolutionary vision of Tim-Berners Lee deeply influenced modern information systems and shaped the construction of the Web as we know it. For instance, the contemporary Internet of Things (IoT) has is roots in Berners-Lee's ideas. Key motivations underpinning the Semantic Web are strongly related to the representation of data. "*The main idea of the Semantic Web is to support a distributed Web at the level of the data rather than at the level of the presentation*" [Allemang and Hendler, 2008]. The seminal document "*Metadata Architecture*" sets the first steps towards the description of web resources adopting metadata, that is "*machine understandable information about web resources or other things*" [Berners-Lee, 1997]. Therefore the Semantic Web can be seen as a great example of the power of representations that unleash functionalities and support the processes of knowledge discovery. These reasons enforce our argument – dedicated and focused experts' effort is required in order to tackle effectively the representation dimension and to fully appreciate its related implications.

The Semantic Web proposed the representation of data as a graph where each resource is connected with its associated meaning. Such representation realises a special type of graph formed by individual triples: $(subject, predicate, object)$. Revisiting an example previously introduced we can illustrate such a triple: "*My name (subject) is (predicate) Luca Trani (object)*". Each element of the triple can be uniquely identified via Universal Resource Identifiers (URIs) or Internationalized Resource Identifiers (IRIs). There is no specific order or predefined hierarchy and any element of the graph can be accessed at any point. Connections, *i.e.* semantic links among resources, can be created dynamically and independently. The resulting model is an open and evolving graph whose vertices and edges capture and maintain information. Knowledge can be inferred by navigating the graph – for instance, axioms can be tested and new ones can be derived. Those are major differences compared for instance to a document-centric model (*e.g.* XML).

Languages and grammars have emerged to formally describe and express representations of the Semantic Web. RDF is one such notation, it is a W3C Recommendation [Manola and Miller, 2004]. RDF allows us to make statements about resources, it can represent the triples of a knowledge graph and can be serialised in a variety of encodings such as RDF/XML, RDF/Turtle, RDF/N3 and JSON-LD. Listing 3.1 illustrates an example of RDF/Turtle.

Listing 3.1: Example of RDF/Turtle serialisation. The namespaces are declared and express the context. It denotes a resource identified by the URI #me and characterised by four relationships.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix  ex: <http://myexample.org/> .
<#me> ex:name "Luca Trani" ;
      ex:familyName "Trani";
      ex:givenName "Luca";
      rdf:type ex:Person .
```

The initial picture of the Semantic Web was successively refined in its implementation leading to the definition of Linked Data [Berners-Lee, 2006; Shadbolt et al., 2006; Bizer et al., 2009]. It was recognised the need to make resources directly accessible by enabling effective navigation of the knowledge graph. The technological solution to support such behaviour was to make identifiers of each resource actionable *e.g.* via hyperlinks. "*The Semantic Web isn't just about putting data on the web. It is about*

making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data" [Berners-Lee, 2006]. To enable such vision Berners Lee indicated four rules to be followed:

1. *Use URIs as names for things*
2. *Use HTTP URIs so that people can look up those names*
3. *When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)*
4. *Include links to other URIs. so that they can discover more things*

In other words the main principle behind Linked Data is to make URIs *de-referenceable* so that both humans and machines can follow them. Linked Data and Linked Open Data (LOD) – its more explicitly open license version – are acknowledged and established realities. Important commercial players manage very large graph-based representations for instance to enhance discovery services *e.g.* Google Knowledge Graph [Uyar and Aliyu, 2015]. Also, application of such representations go beyond data and target devices *e.g.* Graph of Things [Le-Phuoc et al., 2016]. An example of graphical representation of a LOD cloud is provided in Fig. 3.4

Currently the term Semantic Web is broadly used and associated with a collection of specifications, standards, technologies and initiatives supporting the idea of a "Web of Data" [4]. In the following sections we report some of those initiatives as they will be leveraged for our framework described in Chapter 5.

3.2 Organising knowledge

We introduced examples of models for representations and discussed how specific choices, such as graphs, can be particularly suitable to represent relationships among concepts, thereby enabling knowledge discovery. We discussed the use of metadata standards to express such representations – they define element structures, rules and constraints that the corresponding content should fulfil.

In this section we focus on content values and how they can be organised and provided to achieve consistency, shared meaning and mutual understanding. In other

[4] www.w3.org/2013/data/

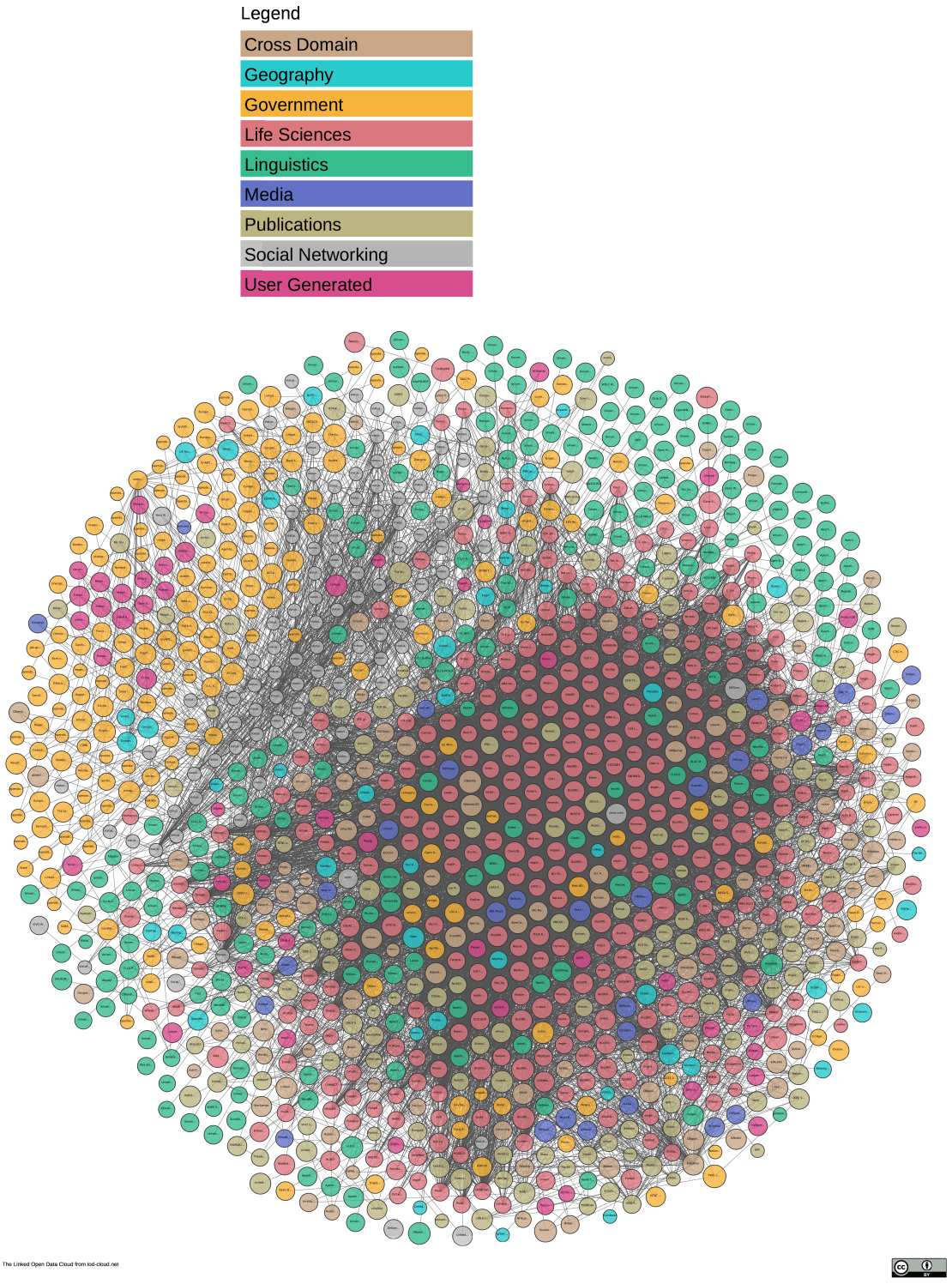


Figure 3.4: Example of Linked Open Data cloud – it represents 1,220 datasets with 16,095 links (as of June 2018)

Source: `lod-cloud.net/`

words here we target the '*value vocabularies*' [Isaac et al., 2011] as opposed to the '*property vocabularies*' (metadata element sets) introduced in Section 3.1.4.

Presumably the simplest way to provide metadata content values is by free text. For example, in the case of metadata describing a book, a field called '*subject*' could be populated with values such as: 'roman history' or 'history of the Romans' or 'the Romans'. Those terms would be all valid and easily interpretable by a human. However, such an unbounded freedom in the choice of values might suffer drawbacks when attempting to enable machine-interpretability. The content provided in this way is highly subjective, prone to errors and ambiguities (*e.g.* due to terms' synonymy and homonymy). Even when syntactical checks are applied, the correct interpretation of the intended meaning might be not guaranteed. To overcome those issues more structure should be applied to content values. This entails socio-technical challenges, *e.g.* to reach agreement on the organisation of concepts and their meanings. Such challenges are extensively discussed in this thesis.

For example, *Term Lists* offer a mechanism to address some of the issues. Instead of working with an open-ended space (*e.g.* free text) they constrain values to an agreed and/or authoritative set of terms or keywords associated with concepts. An extremely powerful approach to organise knowledge adopted since ancient times (*e.g.* by philosophers, naturalists and biologists) are *taxonomies* that group concepts with similar characteristics into categories.

These are examples of Knowledge Organization Systems (KOS) that can be harnessed to provide content values with the sought structure. KOS is a broad term that "*is used in practice to denote systems, tools, and services developed to organize knowledge and to present the organized interpretation of knowledge structures, including automated categorization or knowledge mining software*" [Golub, 2011]. KOS "*encompass all types of schemes for organizing information and promoting knowledge management*" [Hodge, 2000]. They include "*classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names [...] highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies*" [Hodge, 2000]. There exist several descriptions and categorisations of KOS in the literature. Figure 3.5 illustrates one such classification in [Lei Zeng, 2008].

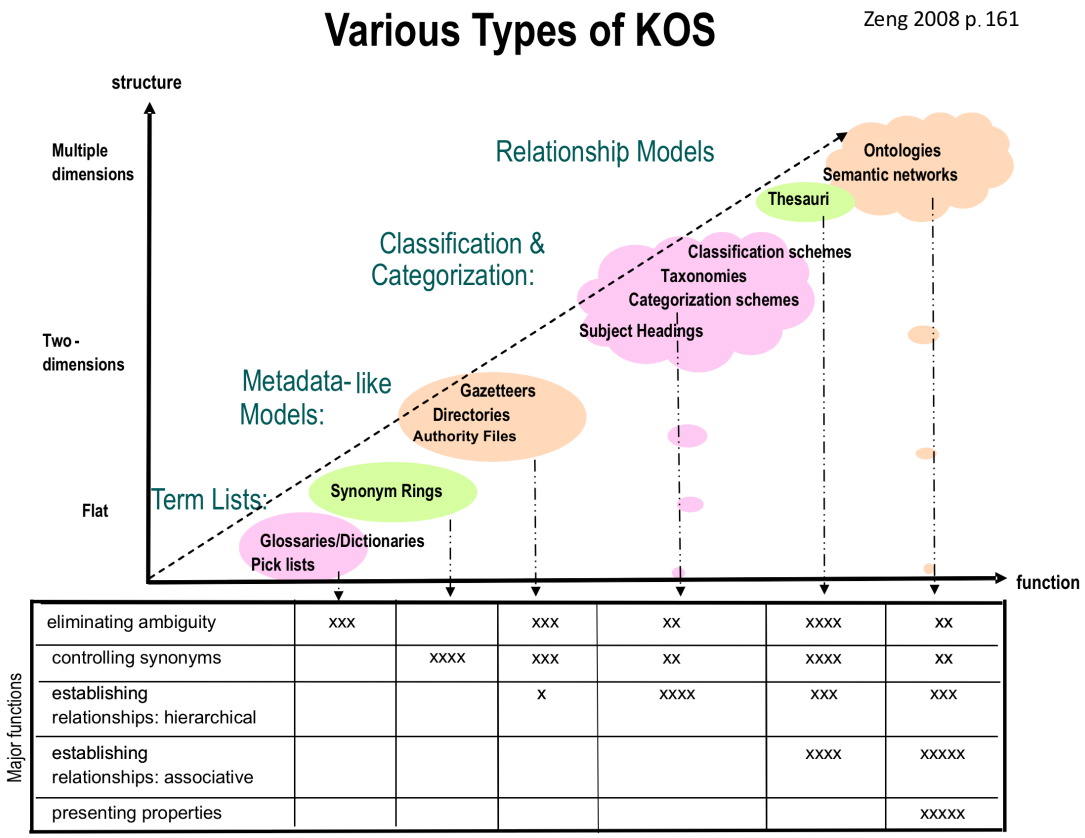


Figure 3.5: A visualisation of KOS arranged by complexity of structure, degree of control and functions enabled. We notice, for instance, how ontologies enable a rich set of functions which is reflected in a more complex structure and formalised controls.

Source: [Lei Zeng, 2008]

An extensive review of KOS is out of the scope of this thesis. We acknowledge the value of KOS as they help establish shared vocabulary and promote agreed meanings. They can be exploited to build, maintain and exchange populations of concepts. For instance, they are extremely powerful used in combination with Linked Data. "*A LOD KOS vocabulary must follow the principles of Linked Data and must be openly available*" [Lei Zeng and Mayr, 2018]. LOD KOS data are described in RDF thus representing populations of organised knowledge that are made available *e.g.* via dedicated vocabulary services. In the following sections we review well-known languages to represent such structures.

3.2.1 RDFS and OWL

In the previous section we have introduced structures that embed relationships of progressive complexity and level of formalisation. To express such structures with their relationships specific languages have been designed. In particular, in the context of the Semantic Web significant efforts have been invested to devise powerful and expressive formalisms. For instance, to enrich the capabilities of RDF a data-modelling vocabulary known as RDF Schema (RDFS) has been conceived. RDFS is an extension of RDF that provides "*mechanisms for describing groups of related resources and the relationships between these resources. [...] These resources are used to determine characteristics of other resources, such as the domains and ranges of properties*" [Brickley and Guha, 2014]. RDFS is widely adopted to define RDF vocabularies. Some major features are:

- Characterising resources into classes with associated properties.
- Modelling structures such as collections.
- Characterising properties with their range and domain.
- Expressing relationships of classes and properties (*e.g.* `type`, `subClassOf`, `subPropertyOf`).

These features make RDFS suitable for the representation of KOS such as controlled vocabularies and taxonomies.

We have introduced ontologies as formal representations that can provide sharable and reusable knowledge – they can be adopted as a means to communicate and share meanings. In Chapter 2 we reported how the process of construction of such valuable assets requires intensive and collaborative effort that can be supported by dedicated methodologies. W3C produced a formal language to represent ontologies, namely the Web Ontology Language (OWL) [McGuinness and van Harmelen, 2004]. OWL is a language designed to represent formal, machine-readable semantics and enable assertion and reasoning. "*An OWL ontology may include descriptions of classes, properties and their instances. Given such an ontology, the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics*" [Smith et al., 2004]. It complements RDF and the RDFS vocabulary by introducing additional constraints and relationships. A set of key features is summarised below:

- Expressing equivalence of classes and properties
- Introducing property relationships such as `inverseOf`, `transitiveProperty` and `symmetricProperty`
- Modelling restrictions on properties of class instances
- Representing cardinalities
- Introducing set operations *e.g.* `unionOf`, `disjointWith` and `intersectionOf`.

OWL is available in different flavours of expressivity, each one targeting diverse communities, uses and requirements: OWL Lite, OWL DL and OWL Full. Details can be found in the official documentation online[5]. OWL 2 introduces syntactical changes and additional features to tackle a broader set of use cases [Golbreich and Wallace, 2012].

A noteworthy consideration about OWL is that it enables representations adopting an *open world* assumption – "*whatever isn't explicitly stated is left as "undefined"—neither true nor false* [Powell and Hopkins, 2015b]. Also, "*descriptions of resources are not confined to a single file or scope [...] New information cannot retract previous information. New information can be contradictory, but facts and entailments can*

[5] `http://www.w3.org/TR/owl-features/`

only be added, never deleted" [Smith et al., 2004]. This is a very powerful feature in that the model deals with incomplete information inherently and it enables extensions. For instance, one could intentionally apply underspecification in the first instance and allow others to reuse by providing specifications depending on their applications. This behaviour is opposed to the *closed world* assumption (*e.g.* 'what is not true is false') underpinning, for instance, XML Schema based representations. Those are less fit to represent knowledge but in turn they offer advantages when dealing with constraining and validating data.

OWL is a very powerful and expressive language but it is quite complex especially for non-experts. Building, representing and maintaining ontologies remains a demanding task. Similarly, measuring their usability and impact can be very challenging [Ma et al., 2018]. In the next section we report about a less expressive and less formal but practical language that gained popularity to build KOS: the Simple Knowledge Organisation System (SKOS).

3.2.2 SKOS

"*The Simple Knowledge Organization System is a common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies. Using SKOS, a knowledge organization system can be expressed as machine-readable data. It can then be exchanged between computer applications and published in a machine-readable format in the Web*" [Isaac and Summers, 2009]. SKOS is an RDF vocabulary that enables the representation of KOS. It is not a "*a formal knowledge representation language*" [Miles and Bechhofer, 2009b], but its data model is represented using an instance of an OWL Full ontology. For example, SKOS most importance resources, *i.e.* `skos:Concept` and `skos:ConceptScheme`, are instances of `owl:Class`. Its main characteristics can be summarised as follows:

- Organisation of *concepts* identified with URIs in *concept schemes*
- Possibility of associating multilingual labels, notations and documentation with concepts
- Representation of semantic relationships between concepts (*e.g.* `broader`, `narrower`, `related`)

- Support for collections (*e.g.* `Collection`, `OrderedCollection`)

These features make SKOS a very powerful representation that in principle can be used also in combination with OWL. The main difference with such a language is that SKOS does not allow us to formally express axioms and facts, therefore it does not enable reasoning and inference. Nevertheless, SKOS offers a powerful data modelling language widely adopted by several communities. An extended version of the SKOS vocabulary called SKOS eXtension for Labels (SKOS-XL), provides a better support for the description and the linking of lexical entities [Miles and Bechhofer, 2009a].

3.2.3 Shapes Constraint Language

The Shapes Constraint Language (SHACL) is a W3C Recommendation that is rapidly gaining interest in the semantic community. SHACL is "*a language for validating RDF graphs against a set of conditions. These conditions are provided as shapes and other constructs expressed in the form of an RDF graph. RDF graphs that are used in this manner are called 'shapes graphs' in SHACL and the RDF graphs that are validated against a shapes graph are called 'data graphs'*" [Knublauch and Kontokostas, 2017]. Shapes graphs are RDF expressions that explain how data is organised. Those expressions include allowed rules, values, patterns and offer a powerful mechanism to formalise constraints and validate data structures. They can be used as templates to model and query data structures. A number of use cases for the application of SHACL are currently under discussion [Steyskal and Coyle, 2017]. The "Open Content Model" (OCM[6]) is an application context of particular interest for us. For instance, according to the OCM multiple independent applications might agree to share the same representation for common data items and allow the presence of undefined data items to account for specialisations in the diverse applications.

Listing 3.2 illustrates an example of a shapes graph, it includes simple constraints but the expressivity of SHACL is much broader. This graph defines the allowed values for a `Person` entity. The data graph defined in Listing 3.1 fulfils the requirements in Listing 3.2 – it passes the validation without raising errors.

[6] `https://www.w3.org/TR/shacl-ucr/#uc24:-open-content-model`

Listing 3.2: Example of application of SHACL. It shows a shapes graph that defines a Person. The data graph in Listing 3.1 passes validation against this shapes graph without errors.

```
@prefix  ex: <http://myexample.org/> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

##A shapes graph that defines a Person
ex:PersonShape
  a sh:NodeShape ;
  sh:targetClass ex:Person ;    # Applies to the class Person
  sh:property [                 #
    sh:path ex:name ;           # constrains the values of ex:name
    sh:maxCount 1 ;             # at most 1 name
    sh:datatype xsd:string ;  # specifies the type as string
  ] ;
  sh:property [                 #
    sh:path ex:familyName ;
    sh:datatype xsd:string ;
  ] ;
  sh:property [                 #
    sh:path ex:givenName ;
    sh:datatype xsd:string ;
  ] ;
  sh:closed true ;   # it's a closed shape
  sh:ignoredProperties ( rdf:type ) . # but it admits an additional property (i.e. rdf:type)
```

In Listing 3.3 we show an example of a data graph inconsistent with this shapes graph and Listing 3.4 shows the corresponding validation results.

Listing 3.3: It shows a data graph that fails the validation against the shapes graph defined in 3.2.

```
## A not valid data graph
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix  ex: <http://myexample.org/> .
<#me>
  ex:name "Luca Trani" ;
  ex:familyName "Trani";
  ex:givenName "Luca";
        ex:name "Giovanni Trani"; ## this will raise an error - I am not allowed to have two names
  rdf:type ex:Person .
```

Listing 3.4: Example of validation results of the data graph in Listing 3.3.

```
### Validation results
[
  a sh:ValidationResult ;
  sh:resultSeverity sh:Violation ;
  sh:sourceConstraintComponent sh:MaxCountConstraintComponent ;
  sh:sourceShape _:n357 ;
  sh:focusNode <urn:x-shacl:dataGraph#me> ;
  sh:resultPath ex:name ;
  sh:resultMessage "More than 1 values" ; ##  error message
] .
```

Figure 3.6 illustrates a snapshot of the SHACL Playground[7], an online application based on the SHACL specifications – it enables users to define shapes graphs and validate data graphs against them.

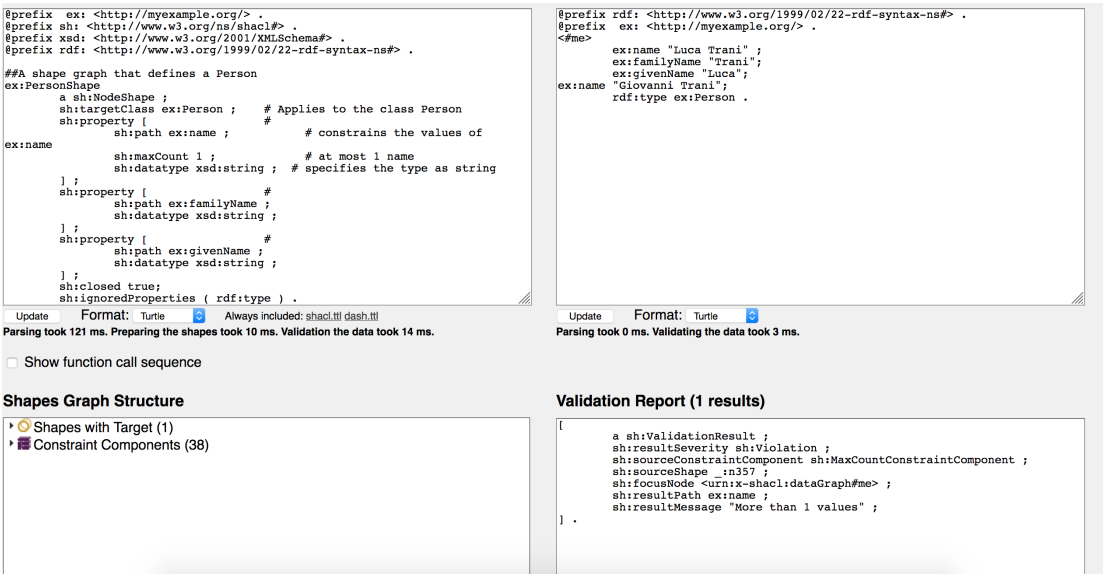


Figure 3.6: A snapshot of an online validator based on SHACL – SHACL Playground.

These examples offer a glimpse of the power of SHACL, such a language will be reprised in Chapter 5 and in Appendix B we present a full-fledged application.

3.3 Examples of metadata standards

In the previous sections we presented languages and formalisms for representations. We now provide examples of popular and widespread metadata standards organised

[7] http://shacl.org/playground/

by their main target applications. These representations are typically the result of collaborative efforts focused on specific community requirements. Later in this chapter we show how such standards can be adopted to build populations by creating concrete instances.

3.3.1 Descriptive metadata

A common use of metadata is the description of a resource. An historical standard for descriptive metadata is the Machine Readable Cataloguing (MARC) developed by Henriette Avram in the 60s while working at the Library of Congress. It was designed "*for the representation and communication of bibliographic and related information in machine-readable form"* and became a standard in 1971 [Avram, 1975]. In the course of the years it evolved introducing new elements and serialisations (*e.g.* XML) and it is still widely used in library cataloguing systems.

The Dublin Core Metadata Initiative (DCMI[8]) produced probably the most popular set of metadata adopted to describe information resources. During a meeting in 1995 a first set of 13 metadata elements were chosen to identify core features of digital objects, the Dublin Core Metadata Element Set (DCMES) [NISO, 2004; Riley, 2017]. That set was successively standardised forming what is today commonly known as Dublin Core (DC). DCMI is an open organisation that is responsible for the maintenance and the evolution of the standard. The set of metadata elements forming the DC vocabulary is intentionally simple thus leading to a wide adoption. It includes: `Contributor`, `Coverage`, `Creator`, `Date`, `Description`, `Format`, `Identifier`, `Language`, `Publisher`, `Relation`, `Rights`, `Source`, `Subject`, `Title` and `Type` [DCMI, 2012]. In the course of the years these elements have been extended and complemented with refinements in an expanded version called DCTerms [DCMI Usage Board, 2012].

3.3.2 Preservation metadata

The OAIS RM addresses the requirements of the digital preservation community. Several metadata standards emerged that are compliant with the OAIS RM. One of the most adopted standards for digital preservation is known as PREMIS Data Dictionary

[8] `http://dublincore.org`

(or briefly PREMIS). It defines preservation metadata as "*the information a repository uses to support the digital preservation process*" [PREMIS Working Group, 2005]. PREMIS stands for PREservation Metadata: Implementation Strategies which is the name of a working group supported by OCLC[9] and the Research Libraries Group, Inc. (RLG[10]) from 2003-2005 that produced the report: *PREMIS Data Dictionary for Preservation Metadata* [Caplan, 2017]. That report defines the standard subsequently published by the Library of Congress as an XML schema. A number of revisions followed and currently the maintenance of that standard is in the charge of the PREMIS Maintenance Activity sponsored by the Library of Congress. Recently the PREMIS Data Dictionary has evolved into an OWL ontology developed by the PREMIS 3.0 Ontology WG in order to support interoperability of digital archives and to facilitate the uptake of Semantic Web technologies in the preservation community [Iorio and Caron, 2016].

3.3.3 Geospatial metadata

A wide spectrum of metadata standards have been produced to represent geospatial resources. In that context organisations such as the Open Geospatial Consortium (OGC[11]) and the International Organisation for Standardization (ISO[12]) have a predominant role in the standardisation processes.

OGC produced several metadata standards and formats to represent for instance: Observation and Measurements[13], Sensors[14], Map[15] and Coverage[16] services and many more.

A well-known family of ISO standards includes:

- ISO19115 for the representation of the geographic information in data.
- ISO19119 for the representation of geographic information of services (*e.g.* geospatial services and catalogues).

[9] www.oclc.org/
[10] http://www.rlg.org/
[11] http://www.opengeospatial.org/
[12] www.iso.org
[13] http://www.opengeospatial.org/standards/om
[14] http://www.opengeospatial.org/standards/sensorml
[15] http://www.opengeospatial.org/standards/wms
[16] http://www.opengeospatial.org/standards/wcs

- ISO19139 XML schema encoding of the geospatial information.

The INSPIRE directives [EU Parliament, 2007] introduced in Chapter 2 include an important part on metadata. INSPIRE leverages a selection of existing standards including Dublin Core, ISO19115, ISO19119 and OGC.

3.3.4 Publication metadata

Another interesting family of metadata standards focuses on the publication of resources in order to make them available and accessible, *e.g.* via catalogues, on the Web. Here we present some examples in different areas. We start with a very successful initiative that targets the publication of Web content: Schema.org[17]. It "*is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond*". Such an activity was initiated in 2011 by Google, Microsoft and Yahoo to describe Web resources and improve the search of content on the Web, thus assisting search engines as they interpret pages in different contexts. Since its conception Schema.org has grown into a popular mechanism to represent structured data on the Web; it is supported by many tools and includes a variety of domains [Guha et al., 2015]. Schema.org is constituted by a hierarchy of classes and relationships and it is compliant with RDF. Typically it is embedded in HTML pages using Microdata [Nevile et al., 2018], JSON-LD [Sporny et al., 2014] and RDFa [Herman et al., 2015].

DataCite[18] is an international organisation founded in 2009 to address scholarly requirements in a wide range of disciplines. Key foundation technology to pursue their goals are persistent identifiers, in particular DataCite endorsed Digital Object Identifiers (DOIs[19]). They also designed an increasingly popular metadata schema, the DataCite Metadata Schema, to represent resources such as scholarly publications and to enable their identification and citation [DataCite Metadata Working Group, 2016].

In the context of our research another important representation is targeting the publication of data catalogues. W3C has invested significant effort steering the development of a vocabulary to facilitate the interoperability of catalogues published

[17] `https://schema.org/`

[18] `www.datacite.org`

[19] `http://www.doi.org/`

on the Web, namely the Data Catalog Vocabulary (DCAT) [Maali and Erickson, 2014]. At present DCAT is a W3C Recommendation that has been endorsed by many players including scientific communities, policy makers and other stakeholders [European Commission, 2017b; Open Knowledge International, 2017]. *"By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation"* [Maali and Erickson, 2014]. Several profiles of DCAT have been produced to address different requirements and there is an active community supporting the uptake of their data model. Furthermore, DCAT is natively supported by catalogue platforms such as CKAN [Open Knowledge Foundation, 2013] presented later in this chapter. Examples of such profiles include: DCAT-AP [European Commission, 2015a] used to describe public sector datasets in Europe; GeoDCAT-AP [European Commission, 2015b] – a DCAT-AP profile describing geospatial datasets, dataset series, and services; and StatDCAT-AP – a DCAT-AP profile for statistical datasets [European Commission, 2016]. One of the key features of DCAT is that it incorporates terms from existing and widely used vocabularies such as Dublin Core, SKOS and FOAF [Brickley and Miller, 2014]. This aspect increases its dissemination and facilitates adoption and uptake into existing systems.

Application profiles try to fill the gaps in the base DCAT standard. Some gaps have been identified and discussed at the "Smart Descriptions & Smarter Vocabularies (SDSVoc[20])" workshop organised by W3C and the VRE4EIC project[21]. Although the current DCAT recommendation is recognised as a powerful tool to improve interoperability of datasets, further work and guidance are needed to extend its adoption and to tailor it to meet community requirements for particular IPC. The W3C Data Exchange Working Group (DXWG[22]) has been recently set up to collect and address requirements from the communities and help improve the DCAT data model. Their efforts yielded a revised version of DCAT which is currently available in a draft version [Beltran et al., 2018] – its model is illustrated in Figure 3.8. In Chapter 5 we introduce

[20] www.w3.org/2016/11/sdsvoc/

[21] www.vre4eic.eu

[22] www.w3.org/2017/dxwg/charter

a representation that we devised drawing on DCAT. It presents commonalities with the model in Fig. 3.8 possibly deriving from mutual influences [W3C-DXWG, 2018; Trani et al., 2018a] – we discuss them in Chapter 6.

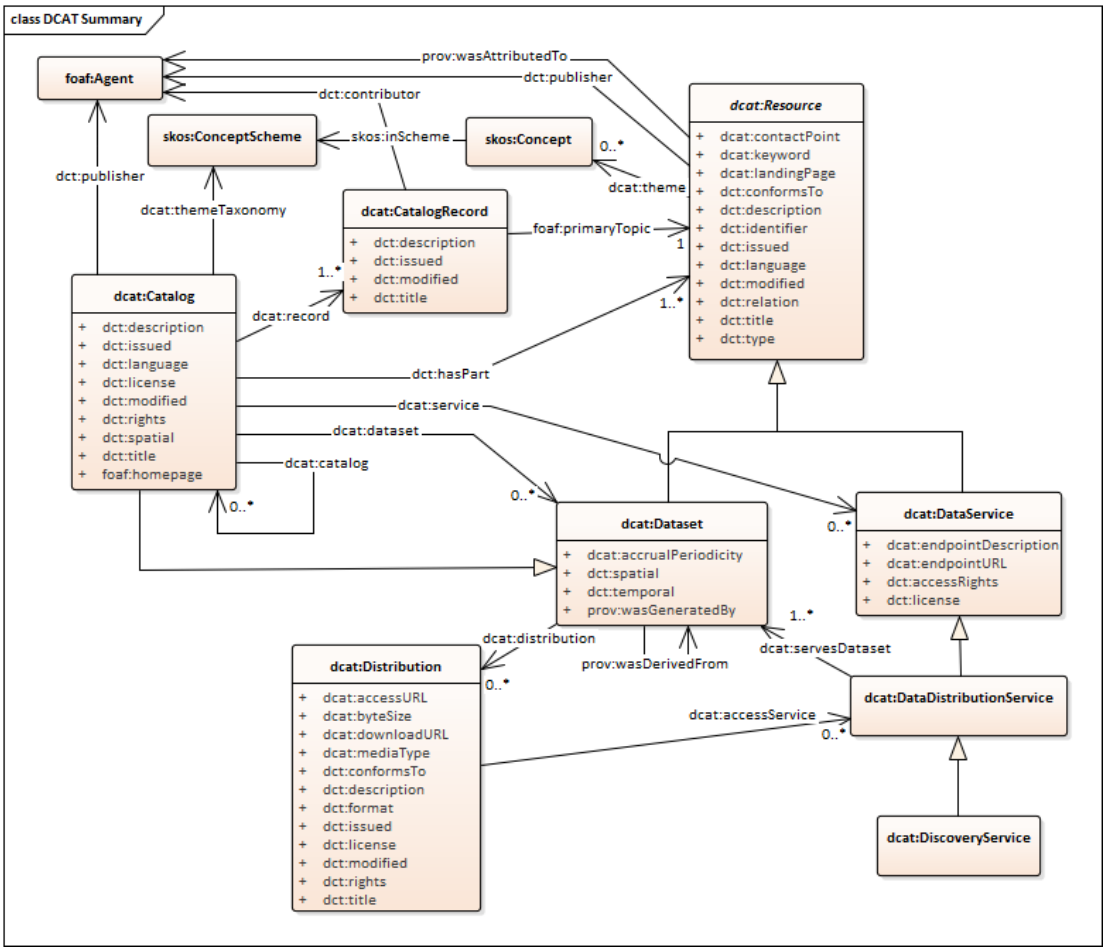


Figure 3.7: An overview of the recently revised version of the DCAT model.

Source: [Beltran et al., 2018]

3.3.5 Metadata and interoperability

We conclude this part of the literature review related to representations by introducing an important application of metadata, namely interoperability. Metadata approaches have been widely discussed as methods to enable interoperability [Veltman, 2001; Chan and Zeng, 2006; Nilsson, 2010; Alemu et al., 2012]. In the context of digital libraries the metadata interoperability issue has been recognised for a long time. As the mission

of digital libraries is to acquire, preserve and provide access to a variety of heterogeneous digital objects, librarians quickly encountered issues related to the appropriate description of digital objects and developed standards and methods for their categorisation. For instance, standards-based metadata, metadata cross-walks or mappings, application profiles and metadata registries have been demonstrated to be valuable methods to enable schema-level metadata interoperability [Haslhofer and Klas, 2010]. Those methods build on a classical interpretation of information organisation systems, mainly hierarchical and authoritative, thus reflecting an objectivist philosophical perspective [Alemu et al., 2012]. However, that perspective has been considered inadequate to organise complex information [Shirky and Clay, 2005]. The advent of social media stimulated collaborative approaches to metadata which exploit social tagging and yield folksonomies [Wal and Thomas, 2007]. Such approaches reflect a social constructivist perspective of the world, they take into account heterogeneous viewpoints, fluidity of interpretation and knowledge sharing [Alemu et al., 2012]. Although authoritative and collaborative, or in other words top-down and bottom-up, approaches might seem antithetic, they can coexist providing complementary perspectives and, as advocated by Gruber, lead to ontology of folksonomy [Gruber, 2007]. Whilst top-down approaches contribute a 'simplified' canonical view according to paradigms of classifications that have been known to humans for a long time, folksonomies recognise the existence of different possible interpretations and account for specialisations and extensions known and understood by subgroups and individuals. The Web Annotation Vocabulary is an example of an ontology supporting such a collaborative approach [Sanderson et al., 2017]. It provides a mechanism to introduce structure and a certain level of formalisation in an area inherently subjective. The support for semantic interoperability and harmonisation does depend on foundations in formalisation sufficient for automation. Arbitrary annotations might require human interpretation thus inhibiting automated methods. The combination of vocabularies for annotations and knowledge organisation systems, such as Controlled Vocabularies, might offer a path to overcome those issues.

Semantic interoperability is a primary goal in this thesis. It entails information sharing and exchange based on negotiated meanings and expressions [Veltman, 2001], it goes beyond the schema-level specifying how metadata records or *content values* are exchanged and used. Therefore, semantic interoperability deals with structure and

includes interpretation leading to mutual understanding of concepts, relationships and their values. Alemu et al. argue that in order to achieve semantic interoperability metadata objects ought to be enriched with knowledge coming from collaborative and user-driven approaches [Alemu et al., 2012]. Semantic Web technologies can provide the appropriate support to achieve semantic interoperability and harmonisation [Nilsson, 2010]. This depends on leveraging declared vocabularies and mechanisms to extend them; unique identifiers that help avoiding naming conflicts and duplications and the ability to express relationships among resources and elements.

3.4 Populations of concepts

In the first part of the chapter we addressed the representation dimension and reviewed methods and approaches to describe and organise concepts. In particular, we discussed the important role of metadata standards and data models. In this section we move to the last aspect of our problem space, namely population. Our focus is on technological approaches that can be leveraged to create and manage instances of concepts, thereby achieving populations. For instance, we review technologies and methods that enable the instantiation of metadata elements and make them available and accessible. In this way populations can be created that reflect established definitions and shared agreements. We address several aspects of the management of populations including persistence, access, retrieval and exchange.

3.4.1 DBMS

Instances of concepts can be created, stored and maintained as entries in databases managed by database management systems (DBMS). DBMS are popular and widespread software systems. A broad spectrum of DBMS exists and they support different underlying data models. A common classification divides DBMS into two categories: relational and NoSQL. The latter was initially coined to refer to category of DBMS that did not make use of the Structured Query Language (SQL) [Lourenço et al., 2015]. However, it then evolved into a variety of systems, *e.g.* based on key-value pairs, arrays, documents or graphs. Our goal here is not to provide a review or comparison of DBMS technologies – for that task we refer to a rich scientific liter-

ature [Angles and Gutierrez, 2008; Loshin, 2013b; Barbierato et al., 2014; Lourenço et al., 2015; Ganesh Chandra, 2015]. We rather list elements that influence the choice of a specific technology depending on use cases and requirements and highlight those that are more suitable in our application scenario. For instance, important aspects to consider are:

1. supported data models;
2. schema definition and query languages;
3. flexibility, extensibility and scalability;
4. supported deployment strategies; and
5. available tooling.

In Chapter 4 we report our experiences evaluating a number of DBMS our selection of a document-based NoSQL DBMS (*i.e.* MongoDB[23]). In Chapter 2 we introduced the data cube which is usually supported by array-based DBMS. Those examples show that the choice of an optimal DBMS for a specific purpose can be driven by qualitative and quantitative considerations.

Recent developments advocate hybrid and multi-model frameworks that are able to address a broad set of requirements by embedding diverse dedicated DBMS. The rationale behind *polystores* [Duggan et al., 2015], as those are also called, is that there is no '*one size fits all*' solution but to achieve better performances applications should exploit collections of specialised engines [Stonebraker et al., 2007]. DBMS typically provide mechanisms to define, manipulate, query and present data via standardised APIs and tools. A common example in the relational DBMS is SQL. In the case of NoSQL the interaction mechanisms might be quite heterogeneous.

In Section 3.1 we discussed graphs as data models particularly suitable to represent and organise concepts and their relationships – they capture and preserve information about data interconnectivity or data topology. Such data models are better supported in DBMS that implement them natively, namely graph databases. "*Graph database models can be defined as those in which data structures for the schema and instances are modeled as graphs or generalizations of them, and data manipulation is expressed*

[23] `http://mongodb.com`

by graph-oriented operations and type constructors" [Angles and Gutierrez, 2008]. They are increasingly applied in a number of scenarios, for instance, to enable big data analytics on large knowledge graphs [Loshin, 2013a], or to connect datasets from open research repositories [Aryani et al., 2016]. Several commercial and open-source implementations exist which differ for features, tooling and support offered [Powell and Hopkins, 2015a]. Also, in some cases the data models might contain slight variations of graph structures – for instance, Neo4j[24] implements a property graph model which includes nodes, relationships and properties (the latter can be associated with nodes or relationships).

In Section 3.1.5 we introduced RDF data models as special types of graphs. Those structures are typically stored and manipulated in so called '*triplestores*'. They are applications that enable the management and retrieval of RDF triples with semantic queries supported by the SPARQL protocol [Prud'hommeaux and Seaborne, 2008]. Popular triplestores were initially built on top of relational DBMS. However, a wide variety of implementations exists including complex systems supporting multiple storage mechanisms, such as Virtuoso[25], that underpins large scale applications like DBpedia [Lehmann et al., 2015]. Recently, native graph semantic DBMS implementations are gaining consensus, example are: Blazegraph[26] and GraphDB[27]. Finally, it is worth to mention valuable advances of research in parallel graph computation that yielded systems such as the GRAPh query Engine (GRAPE) [Fan et al., 2017].

3.4.2 Linked Data frameworks

We have seen that a powerful way to represent concepts and enable knowledge discovery is by adopting the Linked Data paradigm that enforces de-referenceable links associated with each resource. Distributed instances of concepts, *i.e.* populations, can be created and published on the Web by adopting the well-known HTTP protocol. Links, *i.e.* relationships, between resources can be generated independently and dynamically, thereby enabling the evolution of existing populations and/or deriving new ones. A W3C recommendation, the Linked Data Platform (LDP) "*describes the use of HTTP*

[24] https://neo4j.com
[25] https://virtuoso.openlinksw.com/
[26] www.blazegraph.com
[27] http://graphdb.ontotext.com

for accessing, updating, creating and deleting resources from servers that expose their resources as Linked Data" [Speicher et al., 2015]. In this way populations of concepts represented as Linked Data can be created and managed by leveraging HTTP operations such as PUT, DELETE, HEAD, PATCH and OPTIONS. The LDP architecture is based on two main components:

1. LDP clients – HTTP clients that conform to the LDP rules; and
2. LPD servers – HTTP servers that host resources conforming to LDP rules.

The LDP allows servers to manage both RDF and Non-RDF resources, HTTP headers are used to expose information about resources and ways to interact with them. Therefore, LDP supports the Semantic Web vision where resources are inherently distributed and no predefined organisation is imposed besides the few rules presented in Section 3.1.5. LDP has several concrete implementations[28].

A similar initiative that introduces capabilities to modify the values of large graphs of Linked Data (otherwise predominantly read-only) is the Hydra Core Vocabulary – a "*Vocabulary for Hypermedia-Driven Web APIs*" [Lanthaler, 2013, 2014]. It combines principles of REST architectures and Linked Data in order to create evolvable web APIs. "*The basic idea behind Hydra is to provide a vocabulary which enables a server to advertise valid state transitions to a client*" [Lanthaler, 2018]. In this way clients are able to create HTTP requests that, by exploiting the information acquired by servers, modify resources to achieve specific goals. We leverage the Hydra vocabulary to devise a solution that applies in our conceptual framework described in Chapter 5. The approaches presented in this section support the *referencing* behaviour for the Representation Information introduced in Section 3.1.2. In the next section we reprise the OAIS RM to illustrate other relevant viewpoints.

3.4.3 OAIS archives interoperability

In this section we introduce additional aspects of the OAIS RM that help us understand implications related to the deployment strategy and interoperability of distributed archives. As previously mentioned, OAIS concepts focus on preservation. However,

[28] `https://www.w3.org/wiki/LDP_Implementations`

their application can be extended to broader scopes by adapting their design and architectural approaches.

The OAIS Reference Model (RM) recognises the requirement for geographically distributed deployments of OAIS archives – they address the needs of different stakeholders. For instance, consumers may want to see multiple archives as a uniform entity in order to simplify the interaction with them. Managers may be interested in looking for cost-effective solutions which could be achieved by sharing efforts. Producers may want to have a single access point to store their data objects. In order to respond to such requirements archives may wish to cooperate and establish specific agreements.

The OAIS RM identifies four categories of interactions among archives – the first three have a progressive degree of cooperation:

- *Independent* – Archives with no management or technical interactions.
- *Cooperating* – Archives with potential common producers, submission and dissemination standards but no common finding aids.
- *Federated* – Archives targeting both a Local community and a Global community which have interest in the information of different archives provided via one or more common finding aids.
- *Shared resources* – Archives with established agreements for sharing resources. This mode requires the adoption of consistent standards across the archives but does not change the view of the archives presented to the communities.

3.4.3.1 Federated archives

As the target of this research are large federations of independent organisations, more specifically IPC, federated archives are an interesting category that brings in relevant architectural elements for our analysis *e.g.* the potential role of metadata catalogues which are described in Section 3.4.4. OAIS archives may be physically distributed but not necessarily interact with other archives. Therefore, there is a clear distinction between levels of association or interaction and deployment strategies. For instance, an independent archive may be geographically distributed but have a single designated community providing requirements for its preserved information and finding aids.

Figure 3.8 shows a mechanism for realising interoperability among federated archives which serves both local and global communities. In particular, global communities are interested in a *uniform access to federated resources*, thus requiring common finding aids. In this specific example the "*Common Catalog*" addresses the access problem as it provides a "*binding element that serves as a common access point for the information in both Archives*" [CCSDS, 2012]. Local communities preserve their preferred modes of interaction but at the same time global communities gain a coherent and unified view overlooking the federation. "*The Common Catalog may limit its activity to being a finding aid or it may include common dissemination of products from either or both OAISes as shown by the dashed lines in the figure*" [CCSDS, 2012].

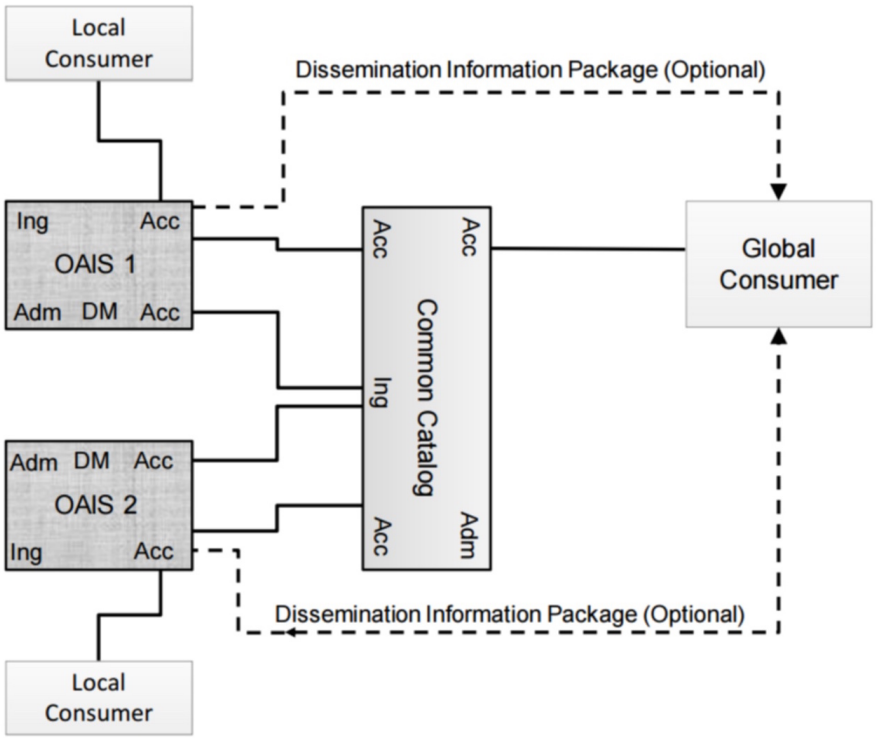


Figure 3.8: A federation of archives that employs a *Common Catalogue* to offer an integrated view of their resources.

Source: [CCSDS, 2012]

Users with enough knowledge about an archive may prefer its specific interfaces, whereas they may query the common catalogue when they need transparent access to federated resources of which they lack specific knowledge. The OAIS RM describes

in much detail the level of functionality of federated archives and classifies them accordingly. It recognises different levels of associations:

1. No association – where there is no interaction

2. Associations that maintain member's autonomy – in this case a member may need to comply to specific requirements to be part of an association but can leave without notice or impact

3. Associations that bind members by contract – in this case to "*change the nature of this association, a member will have to re-negotiate the contract*". Therefore the autonomy is strictly related to the negotiation capability. It is worth noticing that a higher degree of homogeneity may be more easily achieved with more binding contracts.

This example illustrates an architectural approach and its implications to connect existing populations. It exploits metadata catalogues to enable interoperability with a potentially low impact on existing systems in order to preserve the heritage of local communities and their established interaction methods.

3.4.4 Metadata catalogues

In the previous section (3.4.3.1) we have seen how metadata catalogues play an important role in federations of archives. In the context of this research we intend to use such architectural components in a broad sense that might include a variety of features and functionalities beyond cataloguing. Therefore catalogues may become quite complex software systems. Typically they build on top of DBMS, which are exploited as a persistency layer, by adding additional services and interfaces. For instance, they can provide (standardised) APIs for automated discovery and access; administrative tools to manage resources, access and roles; indexing mechanisms; faceted searches; support for multiple metadata schemas; graphical interfaces; data transformation tools; and support for resource publication. In Chapter 4 we present an example of an *ad hoc* metadata catalogue built to tackle the requirements of the seismological community. We discussed the efforts required and the related organisational and technical challenges.

A number of commercial and open-source frameworks exists that might help ease such demanding tasks. One such a framework is the Comprehensive Knowledge Archive Network (CKAN) [Open Knowledge Foundation, 2013]. CKAN is a mature open-source framework that is widely applied in the open-data context, for instance, as backbone for several national open-data portals (*e.g.* Data.gov[29] , Datahub[30]). It targets mainly dataset resources and comes with a rich set of features – it is a comprehensive data management system with an integrated data portal. It offers mechanisms for extension and customisation.

Another example of a popular catalogue framework that was designed for the requirements of the geospatial community is GeoNetwork [Open Source Geospatial Foundation., 2004]. It is a catalogue application "*to manage spatially referenced resources*" which is widely adopted in spatial data infrastructures. GeoNetwork is an open-source software but there are several commercial solutions built on top of it.

Both CKAN and GeoNetwork include features to support distributed deployments and enable exchange and synchronisation of resources. In the next section we provide examples of protocols and standards to enable such features that are often implemented in catalogue frameworks.

In this research we harness metadata catalogues as key architectural components. Rather than providing a comprehensive technological overview our aim here is to help identify critical aspects and challenges. For instance, we recognise that the support for socio-technical, organisational issues, essential for long-term sustainability and longevity, are quite often inadequate in existing frameworks. Lack of specific technical knowledge (*e.g.* necessary for building community extensions and/or for maintenance) might inhibit their uptake and motivate *ad hoc* solutions that on the other hand might suffer from scalability and sustainability problems. A strategic choice should be supported by a cost-benefit analysis addressing aspects such as level and cost of customisations/extensions required and expected to fulfil the targeted use cases.

[29] `www.data.gov/`

[30] `datahub.io/`

3.4.5 Exchanging and synchronising populations

The OAIS RM addresses the interaction modes of distributed archives and the packaging of information (*e.g.* Information Package (IP) that groups data, metadata, contextual information and semantics) without providing technical or implementation details. Those are addressed in dedicated solutions produced by the Open Archives Initiative. For instance, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) "*defines standards for the description and exchange of aggregations of Web resources*" [Lagoze and de Sompel, 2008]. It leverages concepts of the Semantic Web, Linked Data and the HTTP protocol to create de-referenceable aggregation of resources that can be serialised in different formats (*e.g.* JSON-LD and RDF/XML).

One of the most common and widely adopted protocols to enable interoperability of archives is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is "*an application-independent interoperability framework based on metadata harvesting*" [Open Archives Initiative, 2002]. Catalogues implementing OAI-PMH interfaces can expose metadata and process incoming requests, typically generated by OAI-PMH clients (*i.e. harvesters*). Such requests are expressed using the HTTP protocol and can return three types of entities:

- resources – "*A resource is the object or 'stuff' that metadata is 'about'* "
- items – "*An item is a constituent of a repository from which metadata about a resource can be disseminated. That metadata may be disseminated on-the-fly from the associated resource, cross-walked from some canonical form, actually stored in the repository, etc.*"
- records – "*A record is metadata in a specific metadata format. A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item*".

OAI-PMH supports the *collecting* behaviour introduced in Section 3.1.2.

In the geospatial context the OGC CSW[31] was designed to provide a standard interface to catalogue services in order to "*support the ability to publish and search collections of descriptive information (metadata) for data, services, and related information*

[31] `http://www.opengeospatial.org/standards/cat`

objects". CSW has HTTP bindings that enable programs to submit requests leveraging this well-known standard protocol.

Distributed RDF resources can be accessed with federated SPARQL queries [Prud'hommeaux and Buil-Aranda, 2013]. Depending on the targeted scales (*e.g.* number of data sources and sizes) they might entail non-negligible response times and require considerable capacity to achieve acceptable performance. However, optimisation techniques and tools exists to overcome those issues. For instance, SemaGrow is a "*federated SPARQL querying system that uses metadata about the federated data sources in order to optimize query execution*" [Charalambidis et al., 2015].

Resource discovery techniques, such as crawling, and periodical harvesting of descriptions (*i.e.* metadata) might be not sufficient to maintain synchronisation in cooperating systems. In some cases latency and accuracy are critical issues. Typically, an *ad hoc* solution can be implemented to keep alignments of distributed resources. ResourceSync is a standard specification (ANSI/NISO Z39.99-2017) designed for synchronisation frameworks by the Open Archive Initiative. It "*introduces a range of easy to implement capabilities that a server may support in order to enable remote systems to remain more tightly in step with its evolving resources. It also describes how a server should advertise the capabilities it supports. Remote systems may inspect this information to determine how best to remain aligned with the evolving data*" [Lagoze and de Sompel, 2017]. ResourceSync supports synchronisation of both data and metadata. The first might require convenient packaging of resources to achieve cost-efficient data transfers. Packaging standards such as BagIt [Kunze et al., 2018] can ease that task. Similarly, solutions designed for specific use cases are available, *e.g.* for the preservation and distribution of geospatial resources [Pons and Masó, 2016].

The tools and methods introduced in this section show maturity, variety and in some cases broad application. In our research we focus on the identification and definition of requirements that can be adequately supported by existing solutions. As reported in Chapter 5 our envisaged architecture should support the integration of heterogeneous technologies.

3.4.6 Reconciling populations

Exchanging information and knowledge or in other words populations of concepts might lead to inconsistencies. This is a typical scenario when definitions of concepts happen independently and there is no predefined agreement nor authority overseeing them. In Section 3.2.1 we have seen how this situation is inevitable in an open world assumption. Reconciliation techniques can be applied for instance at metadata level to make different representations interoperable [Haslhofer and Klas, 2010].

A very interesting body of literature focuses on ontology reconciliation that is generally a "*human-mediated process, although software can help*" [Hameed et al., 2004]. It builds on the assumption that working with multiple ontologies is the norm as people and organisations naturally tend to use different ontologies. Hameed et al. [2004] identify three main approaches for multiple-ontology architectures:

- Bottom-up – it is based on the principle to "*map between individual ontologies as needed*". There is "*no attempt to identify common, standardised ontologies*". Direct advantages are simplicity and flexibility at the expense of scalability ($O(n^2)$ potential sets of individual bidirecitonal mappings).

- Top-down or common ontology – "*a single common, standard ontology is used as a basis for reconciling the individual ontologies*". There are no individual mappings between ontologies and this allows to cut "*the number of sets of mappings down to just n*". The major effort here is in the construction of the common ontology.

- Clusters of interrelated ontologies – it is a variation of the previous approach. "*Each individual ontology maps to the common ontology for its cluster, and the common ontologies are mapped to allow the exchange of information and knowledge between the clusters*". It combines the advantages of both the previous approaches thus achieving manageability and scalability.

Figure 3.9 illustrates such approaches.

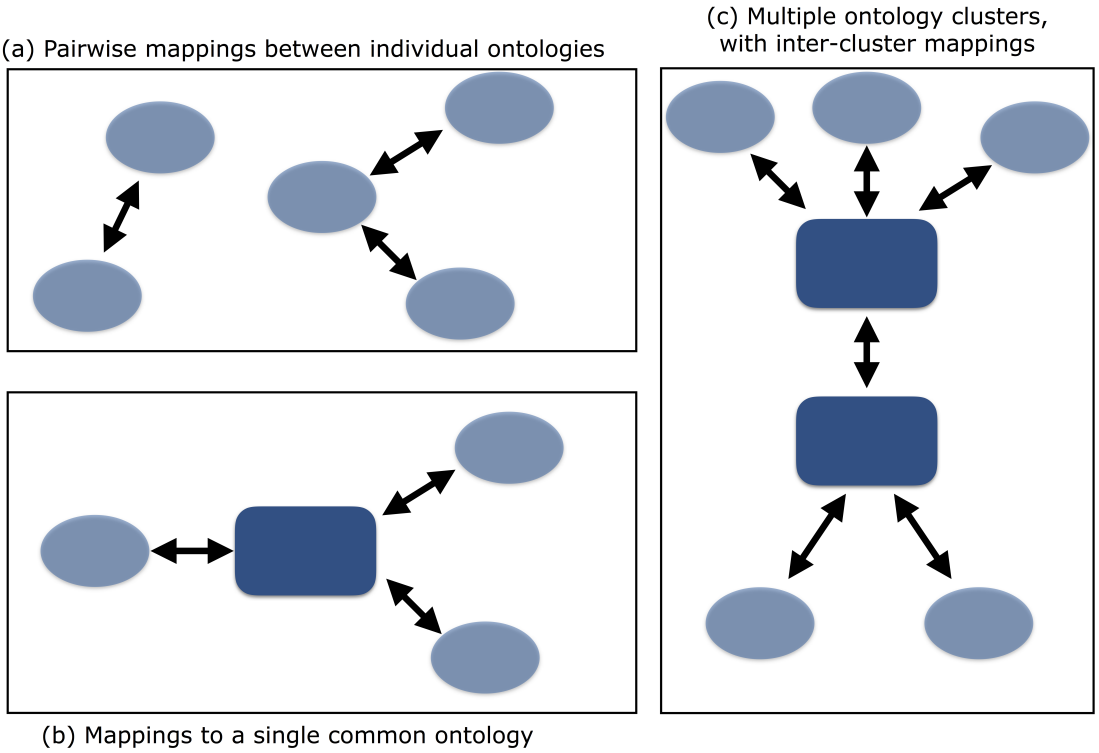


Figure 3.9: Approaches for ontology reconciliation

Source: [Hameed et al., 2004]

Reconciliation is required to resolve *mismatches* between ontologies and can be performed with the following methods:

- Merging – achieved by building a new single coherent ontology that unifies different ones.
- Aligning – involves mutual agreements and shared set of rules to achieve consistency between ontologies that remain separate. It yields alignment statements and relationships between ontologies.
- Integrating – similar to merging but it targets only selected parts of the original ontologies.

In the context of this research we assume that knowledge can be organised and managed in heterogeneous ways. One of our primary goals is to preserve this diversity

characteristic of IPC. Attempting to apply reconciliation on a wide scale in such complex environments is a very demanding task that requires huge human efforts and therefore might be unfeasible to sustain. In Chapter 5 we devise a framework that accounts for such heterogeneity and promotes reconciliation only where strictly needed. Although our focus is not exclusively on ontologies and includes less formal knowledge organisation systems, we acknowledge the great value of the approaches and solutions proposed by the ontology literature and build on them. For instance, we recognise the importance of approaches that assemble hybrid data into RDF graphs [Byrne, 2009]. Also, of particular interest for us are: the approach in Figure 3.9 (c) and the *integrating* method.

3.5 Distributed cross-disciplinary platforms

We conclude our review by presenting examples of systems that combine some of the technologies and approaches reviewed in this chapter in order to integrate information from heterogeneous distributed sources.

RD-Switchboard[32] is a collaborative project conceived and administered by the RDA Data Description Registry Interoperability (DDRI) Working Group. RD-Switchboard "*addresses the problem of cross platform discovery by operating online services that connect datasets across multiple registries*" [Aryani et al., 2016]. It leverages a layered architecture that includes:

- Data Provider Layer – it enables to harvest metadata records via OAI-PMH harvesters.
- RD-Switchboard Inference Components that aggregate, harmonise and store information as graphs.
- Data Access Layer – it enables users to navigate the generated knowledge graphs (nodes and relationships) and makes them available to research communities.

RD-Switchboard adopts the ResearchGraph model to represent Research Data and their connections with Publications, Grants and Researchers. "*By mapping the links between research datasets and related resources, the graph dataset improves both their*

[32] http://www.rd-switchboard.org/

discovery and visibility" [Aryani et al., 2018]. RD-Switchboard is an advanced system that provides integration and harmonisation capabilities. However, it is quite general purpose as it includes many existing aggregators that pull content available from openly accessible sources. A direct downside is that concepts are not chosen or discussed by a community that then takes the responsibility of their maintenance and evolution. Consequently, there are aspects of governance that are lacking, for instance to enable control of the content, relevance and size of the set of concepts.

Another example of advanced data management system is MetaStore that aims to "*provide a generic and reusable metadata management system that can be adopted by multiple scientific communities with different needs*" [Prabhune et al., 2017]. It targets heterogeneous metadata models that are gathered and can be managed in an integrated system. Besides the possibility to annotate metadata it does not offer harmonisation or reconciliation capabilities. MetaStore leverages a multi-model NoSQL DBMS, *i.e.* ArangoDB[33], the SKOS data model for domain specific vocabularies, RDF, SPARQL and OAI-PMH for metadata harvesting. Therefore, it provides us a nice example of multilayered, modular architecture and a valuable tool that fulfils many technical requirements by effectively integrating specialised building blocks.

Also, it is worth to mention that recently a major commercial player (*i.e.* Google) released an online tool specifically targeting scientific datasets. Google Dataset Search[34] "*enables users to find data sets stored across thousands of repositories on the Web*" and leverages open repositories and data models such as Schema.org and DCAT.

Although those systems lack some important characteristics, as they integrate existing resources without a support for negotiating and promoting agreement, they show maturity and effectiveness of technologies and methods. They offer concrete paths that we exploit in the next chapters.

3.6 Summary and conclusions

In this chapter we reported technological solutions exploring two dimensions of our conceptual framework: representation and population. We started with methods

[33] www.arangodb.com

[34] https://toolbox.google.com/datasetsearch

to achieve representations of information and knowledge. We leveraged the OAIS RM to introduce the notion of Representation Information object that captures and formalises the context required to understand digital information. We showed how that concept can be related with metadata as a formalism to express, *i.e.* represent, human and machine actionable information. Also, we reported about systems to organise knowledge, KOS, and introduced some of them. We appreciated the achievements of the Semantic Web community and in particular the concept of Linked Data that enable the representation of large knowledge graphs. The value of publishing scientific data as LOD is widely recognised, and models have been proposed in order to enhance their effectiveness and enable science reproducibility by introducing the concept of Research Objects [Bechhofer et al., 2013].

Likewise, aspects of LD (and LOD) are criticised and debated. Organisation of scientific information is often specialised and optimised to fulfil the requirements of their target communities. A variety of tools and methods exist underpinned by solid mathematical foundations. For instance, they can be applied to create and manipulate multidimensional structures *e.g.* holding time-series and physical models. LD representations might not be as efficient and powerful when applied at such scale. Similar issues apply to query mechanisms.

These considerations suggest us that specialised effort should be retained as migrating existing representations to LD might not be cost-effective. Our view is that granularity plays an essential role in the application of LD. A successful strategy might leverage approaches based on LD and layered architectures – at the higher level LD glue together and integrate heterogeneous contexts whereas the lower levels host specialised representations. We leverage such powerful means in the framework described in Chapter 5.

We then moved to the population dimension and reported about approaches to create, maintain and exchange instances of concepts. We introduced DBMS as persistency layers and showed their importance in broader architectures. The OAIS RM provided us with a terminology and conceptual tools to define distributed systems and their mode of interactions. For instance, it showed us a way to achieve collaboration and interoperability in federated archives via metadata catalogues. Such components build on top of DBMS and offer a number of features, tools and protocols that can be leveraged to publish, exchange and retrieve populations. We presented some of those

exchange mechanisms and highlighted the value of reconciliation. This is necessary to provide a common view of heterogeneous, distributed sources of information as it help resolve mismatches and align concepts. The ontology research provides us with effective strategies.

Finally we reported examples of existing modular systems that integrate several components in order to achieve discovery, retrieval and integration of contents from heterogeneous distributed data sources. The variety of tools and methods and their effective application in real case scenarios suggest to us that the technical aspects are covered by mature and established solutions. In some cases there are several choices and therefore guidance and support are crucial.

Our review is far from being exhaustive, and our aim is to provide directions and guidelines that would help making choices. Table 3.1 summarises some of the key findings of this chapter. They help us draw some preliminary conclusions and offer us a path that will be further exploited in Chapter 5 where we devise our conceptual framework and describe an application in a challenging IPC.

Table 3.1: Summary of literature contributions

Element	Lessons learned	Open issues
OAIS RM	Information and Archives Interaction Models	Focused on preservation and on closed systems (no evolution)
Metadata models	Essential formalism for representation	Require guidance
Semantic Web and Linked Data	Powerful mechanisms to capture knowledge	Choice of granularity, focus, usable and efficient tools
Ontologies	Rich expressivity and formalism, reconciliation methods	Complexity, require expert effort and users' engagement
DBMS	Established technology	General purpose, guidance and optimisations
Metadata catalogues	Variety of solutions	Support and governance
Exchange protocols and tools	Wide offer	Guidance for choice and application

To conclude we highlight the main ingredients that will underpin our strategy discussed in the next chapters:

- Metadata and metadata catalogues to represent information
- KOS and LD to organise and publish knowledge as recognised bundles
- Multi-layer architecture to account for diversity and achieve focused agreement

Chapter 4

Meeting the challenge of establishing shared information

This chapter is an adaptation of the published article "`WFCatalog`*: A catalogue for seismological waveform data" [Trani et al., 2017]*[1]*. Unless otherwise indicated, the content reflects the status at the time of publication, a retrospective and progress are reported in Chapter 6.*

In this chapter we explore critical challenges of establishing shared information that we experienced in a focused application context. We report advances in seismic waveform description and discovery leading to a new seismological service and present the key steps in its design, implementation and adoption. This service, named `WFCatalog`, which stands for waveform catalogue, accommodates features of seismological waveform data. Therefore, it meets the need for seismologists to be able to select waveform data based on seismic waveform features as well as sensor geolocations and temporal specifications. We describe the collaborative design methods and the technical solution showing the central role of seismic feature catalogues in framing the technical and operational delivery of the new service. Also, we provide an overview of the complex environment wherein this endeavour is scoped and the related challenges discussed.

[1]The article was conceived, developed and written by myself. Co-authors helped by discussing and refining ideas, reviewing the manuscript, providing inputs and supporting with the software developments. I was responsible of the final editing.

As multi-disciplinary, multi-organisational and global collaboration is necessary to address today's challenges, canonical representations can provide a focus for collaboration and conceptual tools for agreeing directions. Such collaborations can be fostered and formalised by rallying intellectual effort into the design of novel scientific catalogues and the services that support them. This chapter offers an example of the benefits generated by involving cross-disciplinary skills (*e.g.* data and domain expertise) from the early stages of design, and by sustaining the engagement with the target community throughout the delivery and deployment process.

By being actively engaged in the shaping and development of WFCatalog we observed the interactions among experts and experienced the challenge of establishing the approach and of getting it adopted. This influenced our vision and motivated additional investigations presented later in this thesis. We summarise lessons learned in the conclusions of this chapter. We start by presenting motivation and context underpinning `WFCatalog`.

4.1 Motivation and context

In recent years seismology has experienced a paradigm shift accompanied by major innovations and changes. Seismology has become a data-intensive science where the increasing abundance of data plays a crucial role. This change carries inevitable consequences and affects the way seismologists pursue their research. Seismic network operators, data producers and data centres are equally impacted by this revolution. The role of data centres is changing dramatically, moving from being "simple" data repositories to providers of advanced data services, *e.g.* for data and metadata curation, data exploration and access, analysis and processing. Connection and engagement with user communities has helped steer this transition. The availability of easily accessible data and derived products increases the demand on data centres to provide better and more efficient services for their users. Feedback from user communities influences the design of data centres' technical and organisational architectures.

Our contribution is driven by user demand, existing limitations in current seismic waveform data descriptions and the consequent shortcomings of the paradigms of discovery and access. These limitations provided the motivation for improving the interaction mechanisms between users and seismological data centres.

This chapter presents a novel approach to seismic waveform description which is central to the enhancement of seismic discovery and access services – we describe a technical solution which has been implemented and deployed in the major European seismological data centres federated in the ORFEUS European Integrated Data Archive (EIDA[2]).

4.1.1 Seismological data and access

A typical modern seismic station provides continuous, 3-component recordings of ground motion that are typically between 1 and 100 samples per second. A seismic network comprises a number of geographically distributed seismic stations, from which the data streams usually are transmitted in real-time to a data centre. Here, data are archived, processed and analysed by seismologists to extract seismological information (*e.g.* earthquake location and sub-surface structure).

Seismic waveforms are the "primary" data and the seed that yields a multitude of higher-order derived products, thus they should be treated as first class citizens in seismological data centres. Observatories and Research Facilities for European Seismology[3] is the organisation that coordinates the seismic waveform data acquisition and provisioning in Europe. Under the aegis of ORFEUS, EIDA provides a framework to define and share policies for seismic waveform data acquisition, curation and access.

Refining and improving data services according to users' requirements is a major task for EIDA. It requires a deep understanding of and engagement with the user community. The requirements of this community are continuously evolving, thus presenting new challenges to data and service providers. Methods and data analysis techniques have an impact on data management and contribute to pushing the limits of existing infrastructures. For instance, data intensive techniques, like cross correlation of accumulated datasets [Galea et al., 2013; Addair et al., 2014], require the efficient management of and provisioning for large volumes of data.

Typically an analysis workflow starts with *data acquisition*. This time and resource consuming step entails users' interaction with one or several data centres. Data centres usually offer several methods and tools to support users' data acquisition providing dis-

[2] http://www.orfeus-eu.org/data/eida

[3] www.orfeus-eu.org

covery and access to their data holdings. These tools are continuously improved and have gone through substantial enhancements, for instance moving from email based tools (*e.g.* BreqFAST) to web services (*e.g.* FDSN web services[4]). The latter enable machine-to-machine communication which is a fundamental requirement to achieve automated workflows. Nowadays, many scientific methods in seismology are encapsulated and formalised as workflows, drawing on standard libraries for data handling and transformation [Krischer et al., 2015; Filguiera et al., 2014; Atkinson et al., 2015]. The automatic enactment of such workflows poses additional requirements on the data services, such as managing their rapid bursts of requests for data access, distributing resources and responses according to agreed policy and automatically maintaining usage and accounting records to justify resources and support planning.

The paradigm underpinning the request of seismological waveform data has remained almost identical for many years leveraging well-known and common query parameters – including sensor (*e.g.* network, station and channel) and temporal descriptions (*e.g.* start-time and end-time). This set of parameters is well-known among seismologists and satisfies the requirements of several use cases. Nevertheless, the current data services suffer from important drawbacks *e.g.* the lack of a mechanism to check the availability of a certain dataset, or lack of an overview of the content of seismic streams. In most of the current seismological data services seismic waveforms are treated as *opaque* objects, meaning that very little information is exposed about their actual content. Direct consequences of such shortcomings are:

1. higher rates of unusable data downloads;
2. increased load at users' sites, in terms of data volume and CPU usage; and
3. higher rates of request misses.

Leveraging on users' requirements we reduce some of these shortcomings in the current seismic waveform description, discovery and access methods.

[4] `www.fdsn.org/webservices`

4.2 Methods

At the foundation of this effort there is the concept of a catalogue. This catalogue organises and conveys information embedded in continuous seismic streams, *i.e.*, it is a *seismic waveform feature* catalogue. Catalogues are commonly used in seismology *e.g.* to collect and distribute seismic events [Godey et al., 2013], historical earthquakes information [Albini et al., 2013] and strong motion parameters [Cauzzi et al., 2016]. To the best of our knowledge, the description and discovery of seismic waveforms in terms of their content has not been addressed so far.

Building and populating such a catalogue requires a good understanding and knowledge of the seismologists' practices and the common patterns of seismic data analysis. Without direct access to such features users would have to compute them on datasets downloaded as opaque objects, risking that unwanted characteristics would lead to data disposal. Potentially this situation may result in a vicious circle with a conspicuous waste of resources. Instead we pursue a *virtuous circle* with an efficient use of resources which is a fundamental requirement of any data-driven science. The key to invert this cycle has been identifying a number of tasks and operations of general concern and moving them from *users' sites* to *data centres' sites*. Moving repeated resource consuming tasks into data centres provides several advantages:

1. it reduces users' resource consumption supporting more efficient use of resources at data centres;

2. it leads to a *canonical* definition and representation of seismic waveform features; and

3. it supports and enhances data discovery and access services making them tailored to users' requirements.

For instance, a data centre can cache the results of common operations, thus amortising the computational costs over many users. Also, data centres can tune and optimise the performance of such computations and develop the necessary expertise. This can be seen as a *delegation of responsibilities* from the users to the data centres that must deliver: **trust** and **reliability**, and provide **verifiable** and **guaranteed results**.

In the subsequent sections we present details of the `WFCatalog`'s operations, data model and architecture.

4.2.1 `WFCatalog` operations

`WFCatalog` supports several operations:

1. computation, collection, ingestion of metadata
2. stewardship of metadata – update, delete, versioning
3. query functionalities
4. metadata publication
5. data access based on queries over the metadata

Metadata computation, collection and ingestion (1) are core functionalities provided by `WFCatalog`. The computation of metadata is performed close to the related data archive, and requires direct access to the raw seismological data. Computation can be scheduled according to a configurable frequency – this feature provides flexibility and allows us to meet the related policies within the federation. The management of metadata (2) must reflect the agreed policies and the data lifecycle.

`WFCatalog` provides *readonly* capabilities to its users. Metadata ingestion and update are delegated to data centres' operators. This choice reflects the idea that data centres are responsible for the curation of their data holdings, which includes the generation and curation of the related metadata. Metadata may have different versions identified by a timestamp and a version number. At present querying the catalogue for specific metadata versions or performing timestamped queries is not supported – the most recent version of the metadata is provided by default. This behaviour will be extended in future releases in order to facilitate reproducibility. Different query patterns (3) are currently implemented. *Multifaceted* queries spanning across multiple parameters are supported, including: temporal constraints, stream specifications (network, station, channel, location id, *etc.*), quality parameters and continuous segments (see Table 4.1). *Multisite* queries are not supported because data centres should expose only the information, data and data products which they are responsible for. Metadata publication (4) is essential when considering cross-disciplinary science. Adopting standards to publish datasets enables easier discovery and interoperation in broader contexts. `WFCatalog` supports the usage of *Persistent*

Parameter	Type	Description
network	string	network code
station	string	station code
channel	string	channel code
location	string	channel location identifier
starttime	ISO8601	start time of the selection
endtime	ISO8601	end time of the selection
format	string	specify the output format (default JSON)
include	string	specify the level of detail of the results, *e.g.* $include = sample$
granularity	string	define the desired level of granularity, *e.g.* day
minimumlength	float	limit results to continuous data segments of a specified minimum length in seconds
longestonly	boolean	limit results to the longest continuous segment per channel
csegments	boolean	include information about continuous segments
[metric_filter]	metric dependant	select streams that satisfy a filter on a specific metric value for any metric defined in table 4.2, *e.g.* $sample\_max\_lt = 10$ & $sample\_max\_gt = 3$

Table 4.1: Query parameters supported by `WFCatalog` (October 2018)

Identifiers, which entails a commitment to guarantee access to metadata even beyond the data lifespan. `WFCatalog` improves discovery and access (5) to waveform data. At present direct access to data is not provided, but it can be enabled in combination with data access services *e.g.* `fdsnws-dataselect`[5]. The partial API compatibility allows the sharing of queries across services. In a future release persistent identifiers pointing to the data objects (*e.g.* EPIC Handle[6]) will be embedded in the responses from `WFCatalog`.

[5] `www.fdsn.org/webservices`

[6] `www.pidconsortium.eu`

4.2.2 Data model

Improving the description and representation of seismic waveform is a major goal of this effort. Such a representation should be:

1. recognised and shared;
2. flexible and extensible; and
3. lightweight and suitable for machine-to-machine communication.

The interoperation with broader multidisciplinary environments demands clear formats and well-defined interfaces. Our solution has been designed with these general requirements in mind; we also address *attribution*, *citation* and *reproducibility*.

4.2.2.1 Data quality metrics

We perform the qualification of seismic waveform according to well-defined and agreed *data quality* metrics, which can be derived from seismic waveforms. The selection of the metrics is not a trivial task and it has been accomplished in successive steps involving several stakeholders. Besides the purely technical issues there are other relevant aspects to consider. A major hurdle is the difficulty to find a common, meaningful and shared way to define *data quality*. The interpretation of data quality is often subjective and varies significantly from case to case. Metrics should span a broad set of use cases and target different users. The selection process was initiated and carried out in the context of EU FP7 project NERA[7] [Sleeman, 2014a,b]. This delivered a coherent preliminary set of data quality metrics, which have been further developed and endorsed by EIDA data centres. Subsequently, a broader community has been involved by targeting the International Federation of Digital Seismograph Networks (FDSN). That discussion is currently (October 2018) ongoing and a core set of metrics and their associated definitions has been identified. Consensus and shared definitions of such metrics are fundamental requirements to ensure compatibility, exchange and comparison of results across different systems. The list of data quality metrics adopted in the current version of `WFCatalog` is provided in Table 4.2. For a more complete overview we refer to the `WFCatalog` specification [Trani et al., 2016].

[7] `www.nera-eu.org`

Sample metrics	
num_samples	sample_max
sample_min	sample_mean
sample_median	sample_stdev
sample_rms	sample_lower_quartile
sample_upper_quartile	num_gaps
num_overlaps	max_gap
max_overlap	sum_gaps
sum_overlaps	percent_availability
MiniSEED header metrics	
encoding	num_records
quality	record_length
sample_rate	timing_correction
timing_quality_mean	timing_quality_median
timing_quality_lower_quartile	timing_quality_upper_quartile
timing_quality_max	timing_quality_min
data_quality_flags	
amplifier_saturation	digitizer_clipping
spikes	glitches
missing_padded_data	telemetry_sync_error
digital_filter_charging	suspect_time_tag
activity_flags	
calibration_signal	time_correction_applied
event_begin	event_end
positive_leap	negative_leap
event_in_progress	
io_and_clock_flags	
station_volume	long_record_read
short_record_read	start_time_series
end_time_series	clock_locked

Table 4.2: Data quality metrics implemented in `WFCatalog` (October 2018)

Feature name	Description
wfmetadata_id	identifier of the returned metadata document
producer: data centre, agent, date creation	producer of the metadata document
waveform_type	type of the related waveform, *e.g.* seismic and infrasound
waveform_format	format of the related waveform, *e.g.* MiniSEED
version	progressive number indicating the document version

Table 4.3: `WFCatalog` additional features

4.2.2.2 WFMetadata schema

`WFCatalog` publishes metadata according to the Waveform Metadata (`WFMetadata`) JSON schema [Trani et al., 2016]. This schema is novel and represents (seismic) waveform metadata including data quality metrics and additional features as shown in Table 4.3. An important feature is the possibility to extend its applicability beyond seismological waveforms *e.g.* infrasound time-series data. The `WFMetadata` schema sets the basis to become a standard way to represent and exchange seismic waveform metadata, thus filling a gap in the current seismological metadata offerings. Noteworthy is the support for *Persistent Identifiers* which coupled with versioning and information about the producer, foster proper attribution, citation and reproducibility.

4.2.3 Architecture

`WFCatalog`'s architecture is modular and is composed of the following main elements: data analysis and metrics computation module, metadata store and web service API (Fig. 4.1).

4.2.3.1 Data Analysis and Metadata Computation module (DAMC)

Waveform data analysis and data quality metrics computation are core functions. They yield the features extracted from waveform data which are then stored and made accessible to users. The DAMC implements these operations complying with specified and agreed metric definitions. Our strategy has been to decouple the definition of the features from their implementation. As a consequence, each data centre has the freedom to implement their own DAMCs as long as they comply with the agreed

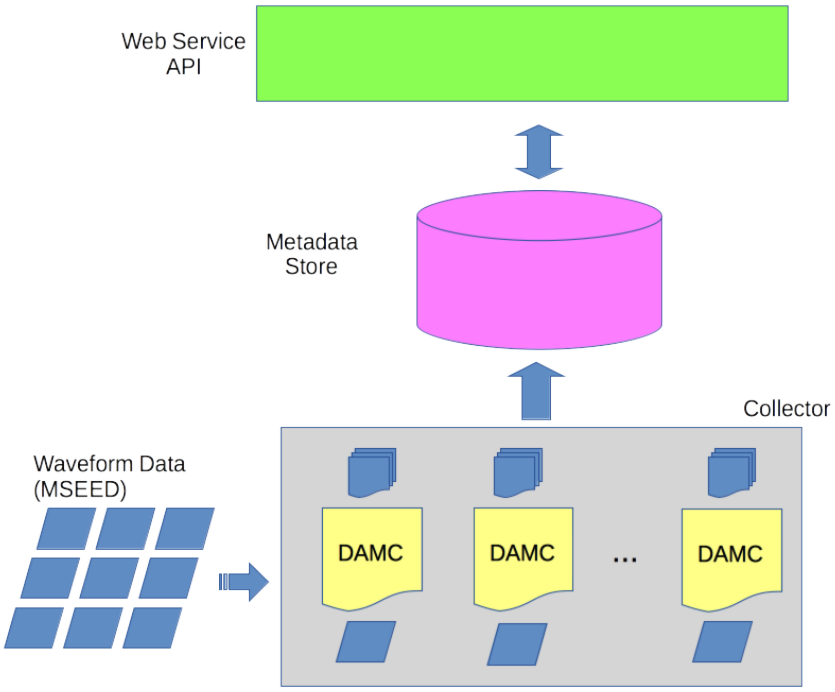


Figure 4.1: `WFCatalog` architecture overview – Seismic streams encoded in MiniSEED feed the `Collector`. This component performs parallel processing of seismic streams instantiating multiple Data Analysis and Metadata Computation modules (DAMCs). DAMC's implementation builds on a popular seismological software library: ObsPy [The ObsPy Development Team, 2016]. Each DAMC extracts features and metadata from seismic streams and populates the Metadata Store. The `WFCatalog` web service API provides programmatic access to the metadata stored in the database.

definitions. We provide a reference implementation which is adopted across EIDA data centres. This implementation builds on top of a popular community-driven Python library, namely ObsPy [Krischer et al., 2015], and it takes MiniSEED (a subset of SEED [Ahern et al., 2009]) data as input. We chose MiniSEED as it is by far the most used format in EIDA data centres and for compatibility with the `fdsnws-dataselect` web service. However, `WFCatalog` and its metadata model are not bound to any specific data format. Liaising with the ObsPy developers and the user community we developed extensions and made an additional module available[8]. The inclusion of the DAMC

[8] `http://docs.obspy.org/packages/autogen/obspy.signal.quality_control.html#module-obspy.signal.quality_control`

code in a software library widely used by seismologists has been a strategic choice with several advantages:

1. it establishes a direct communication channel with the user community;
2. it involves the user community in its design, maintenance and evolution;
3. it enables users to have the same functionalities available at their sites; and
4. it builds consensus and promotes adoption.

The DAMC is configurable and integrated in the `WFCatalog` ingestion process, namely `WFCatalog Collector`. At this stage features are computed with a *daily* granularity. The DAMC, and the `WFCatalog`, have been designed to scale in terms of new features and/or additional time granularities. The ingestion into the database is performed by running multiple DAMC processes in parallel managed by the `Collector`. We store hash signatures which can be used to trigger re-computation of the features when changes occur in the source data files[9].

4.2.3.2 Metadata Store

The features extracted from seismic waveform data require the support of a suitable database infrastructure offering: scalability, performance and optimisation in terms of storage space, query functionality and query response time and costs. We decided to benchmark several technologies before opting for a solution. The final choice has been driven by pragmatic aspects, and it might evolve over time as the architecture is technology independent. Our evaluation considered the following factors: maturity, language support, availability of connectors and software libraries, scalability and extensibility. A complete technology review is out of our scope, we addressed this topic also in Chapter 3. The systems we evaluated were: MySQL, MonetDB [MonetDB BV, 2013], Cassandra [Apache Software Foundation, 2013a], CouchDB [Apache Software Foundation, 2013b] and MongoDB [MongoDB, Inc., 2016]. Of particular interest was the experience with MonetDB [Ivanova et al., 2013a,b]. This technology, when further developed and refined, has the potential to provide functionalities hardly

[9]For instance, updates might be necessary to include delayed data packets from the originating sensors *e.g.* due to transmission issues

achievable with the other candidates. However, at the time when we were selecting our database this technology was not considered stable enough for production and we opted for another DBMS: MongoDB. MongoDB is a very popular document store which provides native scalability and its internal model is flexible and allows for extensions.

In the current setting the database hosts two collections: one holds the features computed on daily files whereas the other holds the features about the continuous segments contained in a specific day with start-time and end-time of each segment. Therefore, we are able to provide a detailed description of the availability of data in each waveform stream. Moreover, the loose coupling of the collections allows for extensions that include additional features and time granularities, *e.g.* hourly.

4.2.3.3 Web API

The web API facilitates the interaction with third-party software and users. This component has to promote usability, support diverse use cases, address the evolving nature of the user community's requirements and allow for extensibility – ideally it should be possible to add features and modify the current query patterns according to new scientific methods whilst maintaining backward compatibility.

The design of the API of the `WFCatalog` has been an iterative, collaborative work involving several stakeholders including data-centre operators, developers and seismologists. The participation of several actors from the early stages of the design contributed useful perspectives and requirements.

The discussion was triggered by a prototype showing the potential capabilities, this prototype has been refined incrementally during further stages. One of the requirements was to allow compatibility with existing service standards (*e.g.* FDSN WS[10]). This reduces the learning curve and facilitates the uptake of a new service. It also enables users and data curators to retain the value of prior investments in methods, workflows, code and working practices. We were able to fulfil this backward compatibility constraint only partially as we replicated methods and some of the query parameters.

Table 4.4 summarises the available methods, for an extensive description we refer to the published web API[11].

[10] `www.fdsn.org/webservices/`

[11] `https://www.orfeus-eu.org/data/eida/webservices/wfcatalog/`

Method	Description
query	enables metadata queries with the supported parameters, returns results in the requested format
version	returns the version of the web service
application.wadl	returns the WADL document describing the service

Table 4.4: `WFCatalog` webservice API methods

4.3 Challenges

In the previous sections we described how we implemented the `WFCatalog`, interpreting users' requirements and translating them into a concrete architecture. In the next sections we introduce the challenges encountered during the design and construction process. Recognising the main challenges and their implications can provide a better understanding of the complexity of the environment in which this work is framed. These challenges can be divided in two sub-categories: *socio-political* and *technical*.

4.3.1 Socio-political challenges

Seismology has a long tradition of global collaboration and data sharing, as well as knowledge and experience about definition, design and implementation of data models, formats, services and tools. Consequently the maturity of the community is reflected and formalised in a number of international coordination and collaboration frameworks at global and European scale *e.g.* IASPEI[12], FDSN[13], ESC[14], ORFEUS[15]. The role of such organisational bodies is fundamental to guarantee authoritativeness, trust, acceptance and adoption regarding the form of shared services and the data they deliver. Alongside the official formalised contexts, there often exist community-driven efforts, which may have an equally large impact. These initiatives can be powerful and can offer direct vehicles to reach out to large and broad communities outside the formal schemes. Identifying the key players and stakeholders of a specific community is essential when designing innovative services for such a community. We addressed

[12] `www.iaspei.org`
[13] `www.fdsn.org`
[14] `www.esc-web.org`
[15] `www.orfeus-eu.org`

a mix of official and *de facto* processes in order to facilitate the definition and uptake of the `WFCatalog`. We targeted FDSN, ORFEUS and EIDA as formal frameworks and the ObsPy as community-driven effort.

The European seismological landscape has a distributed organisation with responsibilities shared across a number of recognised data centres. This organisation has historical and cultural roots but it is also a design choice to address the evolving data challenges. An example is the official establishment of EIDA within ORFEUS in 2013. Previously the ORFEUS Data Centre (ODC) was the centralised European data archive. The newly constituted federated structure responds better to modern challenges but it requires well-defined, shared agreements and a common vision.

Another important aspect is understanding the users, their requirements, the set of tools and methods they use and the limitations of these tools. They are continuously evolving, driven by new scientific insights and technological opportunities – it is important to recognise and respond to such changes. In the seismological domain there are a number of well-known and widespread tools, libraries, methods and data exchange standards. Seismologists exploit such common building blocks by applying customisations and extending them with new methods. However, the sharing of methods and algorithms for data analysis and processing, in the form of workflows, is quite new to the community and gained popularity only recently supported by initiatives such as the VERCE project [Atkinson et al., 2015]. Customisation may inhibit the adoption of standard representations.

Seismologists are accustomed to delegating data management operations to data centres whereas processing and analysis remain the users' focus. A reason for this may be the novelty of method sharing and the feeling that they lack "control" when delegating operations.

An important lesson learned is that technical changes ought to be supported and sustained by appropriate organisational frameworks. These frameworks can provide the context and vocabulary to steer collaborative discussions, pooling insights and efforts thereby accelerating convergence on solutions. They offer trusted environments that facilitate technology uptake and long-term sustainability. `WFCatalog` is the result of a collaborative work initiated within EIDA that provided a proper organisational framework to exchange ideas, requirements and define strategies and policies. These elements are equally important because a catalogue is not just a piece of software – a

fundamental component is related to the authoritativeness of the information therein maintained and offered to the users. Clear and well-defined policies to building, operating, revising and decommissioning such catalogues are key elements.

The combination of policies, software and communication providing high compatibility across a federation like EIDA, allowed us to reach the highest level possible of agreement among partners, whereas such consistency was not replicated at the FDSN. FDSN provides a broad platform to coordinate, discuss, promote and exchange ideas, nevertheless, it has a looser coupling among participants which is reflected in a slower pace to forming global agreements. An additional factor in the agreement forming is the level of commitment which can vary depending on priorities, available resources, *etc*. In our case the clear engagement of most of the contributing partners in an overarching European strategic research infrastructure for solid-Earth science, namely the European Plate Observing System (EPOS[16]), constituted an accelerating factor. Inevitably the boundary conditions provided by EPOS influenced the timeline ensuring a rapid convergence towards a common goal. Therefore the catalysing role of projects and research infrastructures should not be underestimated.

Within large collaborations a major challenge is the multiplicity of factors that need to be synchronised and aligned for a common purpose. Communication, engagement and commitment to key roles by representatives are essential. In order to foster these activities technical architectures need to reflect the complexity of the surrounding environment and offer intellectual ramps [Atkinson et al., 2010].

4.3.2 Technical challenges

The computation of the quality metrics presented several challenges. In order to align the theoretical definitions with the computation, we had to overcome several issues mainly introduced by the SEED [Ahern et al., 2009] data format and by the data archival system. Adaptations allowed us to obtain thorough and accurate results conforming with the definitions. As the system has to cope with the steady growth of the data and the consequent increase of the metadata volume, scalability is essential. The approach and the technology adopted enable us to deal effectively with these issues.

[16] www.epos-ip.org

The chosen data model guarantees the flexibility and extensibility required to address the expansion of the set of metrics and features. Another critical aspect concerning the metrics is the granularity, that is the time window over which the metrics are calculated. As previously mentioned we chose to compute the metrics on daily intervals. This choice is a tradeoff between meaningfulness and pragmatism. The alternatives are fixed time granularity of a different length and dynamic computation tailored on users' requests. The latter represents the ideal solution. For instance, scientists performing analysis on long period signals might be interested in metrics computed and aggregated on a yearly basis. Unfortunately the dynamic solution is also the most expensive from the computational point of view. Also, it requires proper technological support not easily achievable with most DBMSs. We experimented with this approach with MonetDB but for the reasons previously mentioned, maturity and support, we decided to move towards a less advanced but more stable solution. We adopted a fixed granularity but as a mitigating factor we designed the system to accommodate multiple independent granularities.

Another challenge regards the performance of the metrics computation. This aspect influences the database update policies. Ideally a user may want the metrics and the raw data available simultaneously which means near real time. However, processing in near real time the incoming data of thousands of waveform streams can be expensive, especially considering the limited capacity of some data centres. We optimised the DAMC in order to speed up the critical operations. As ObsPy is predominantly a Python framework, in an initial phase some operations proved to be slow and we switched to native C implementation for the critical methods in order to achieve better results. This optimisation provided a gain of a factor of 10 on the most compute-intensive methods.

4.4 Related work

The design, development and deployment of `WFCatalog` was influenced by many aspects of contemporary research including attempts to assess seismic waveform data quality.

Data quality has been a debated topic in seismology for a long time. A number of tools and software packages have been produced addressing data quality. An example

of such software is PQLX [McNamara and Boaz, 2006] which provides a graphical user interface on top of a MySQL database containing probability density functions (PDF) of power spectral densities (PSD). PQLX has been widely used and it became a *de facto* standard for certain metrics.

Another application is the Data Quality Analyzer (DQA) [Ringler et al., 2015]. That application, developed at USGS, follows an approach similar to `WFCatalog` to present data quality metrics and facilitate the assessment of the quality of seismic stations. However, DQA's approach is focused primarily on stations whereas `WFCatalog` is waveform-data centric. Although DQA computes and stores data quality metrics in a central PostgreSQL DBMS it does not expose them as metadata, thus not enabling machine-to-machine communication. DQA has a rich web interface with diverse visualisations.

The above products focus purely on data quality metrics mainly addressing the diagnostics of seismological stations. They do not aim at extending the description of waveform data and they do not provide such metrics as a service. They are valuable tools in the context of their application but they do not address the broader scope. Moreover each solution adopts a different set of metrics and definitions.

The Modular Utility for Statistical Knowledge Gathering (MUSTANG[17]) has a web service interface providing programmatic access to a number of quality metrics in different formats and a data browser for visualising such metrics. Our work is an attempt to homogenise the metrics across different systems. In particular we identified a set of metrics shared with MUSTANG as a basis for discussions at FDSN.

An extensive literature review of related work of this thesis is provided in Chapters 2 and 3.

4.5 Results and discussion

`WFCatalog` is used to assess seismic waveform data quality. It provides data centres with a powerful tool to evaluate and present the quality of their data holdings. Similarly, network operators have an effective instrument that offers them immediate feedback about the status of their sensors, thus helping them address potential issues and delivering better quality data. By offering a catalogue which contains metadata

[17] `service.iris.edu/mustang`

that users would otherwise compute on downloaded data, `WFCatalog` provides major savings:

1. overall computation is reduced because computation results are reused;
2. access to primary data that then proves unusable is avoided, with a substantial network traffic reduction;
3. users do not need to perform quality analyses before they use the data but they still can if they have additional criteria; and
4. gradually the standards for data quality will emerge leading to more consistent science.

The benefits of `WFCatalog` are its data model, its exchange format, `WFMetadata` schema, and the programmatic access to a standardised set of predefined features. `WFMetadata` schema provides a canonical representation of waveform data metadata which helps establish trustworthy communication in a federated environment. The `WFCatalog` constitutes an important addition which combines with other components and services to provide substantial advantages in seismic waveform discovery and access. It helps to steer the discovery process filtering the results tailored by user's requirements about waveform data content. An example of such interaction and service composition is the combination of `WFCatalog` (discovery) and `fdsnws-dataselect` (access).

Service composition is one possible application of `WFCatalog`, another application is visualisation. At the ORFEUS Data Centre (ODC) we developed web interfaces fed by `WFCatalog` in order to check the availability of datasets and visualise multiple data quality metrics. Fig. 4.2 illustrates an example of a visual interface that can be enabled on top of `WFCatalog`. This interface provides a visual inspection of the available seismic waveform streams with a detailed overview of the continuous segments contained in a daily stream.

Figure 4.3 shows another interface which can be used to browse graphically through data quality metrics. This figure shows three examples of metrics computed for different days. This interactive tool allows seismologists to spot possible issues with the underlying data – by clicking on a specific point it is possible to drill down to a preview of the underlying data.

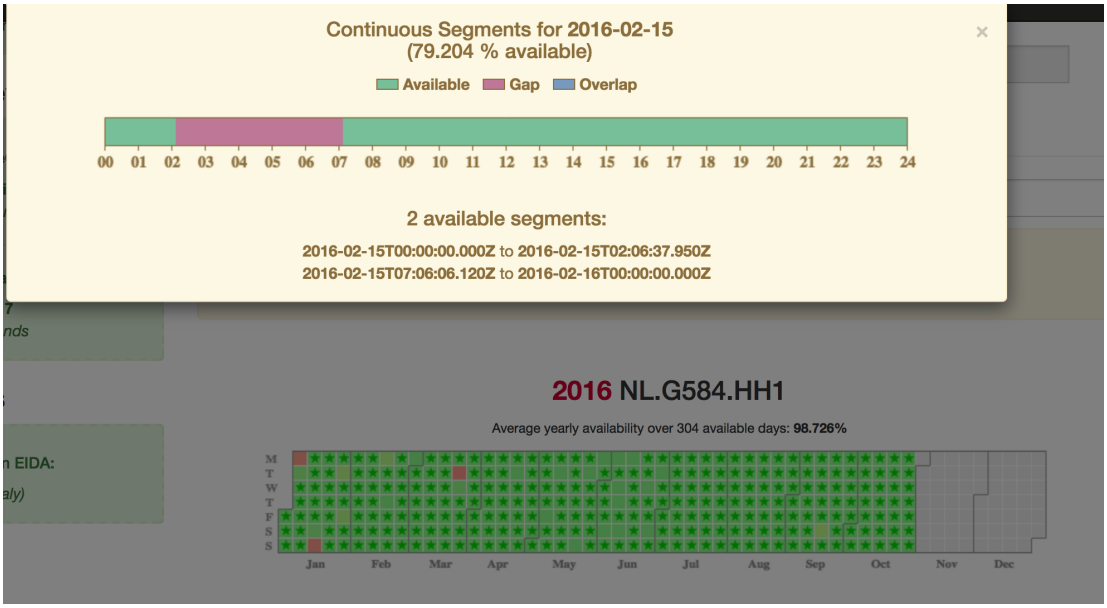
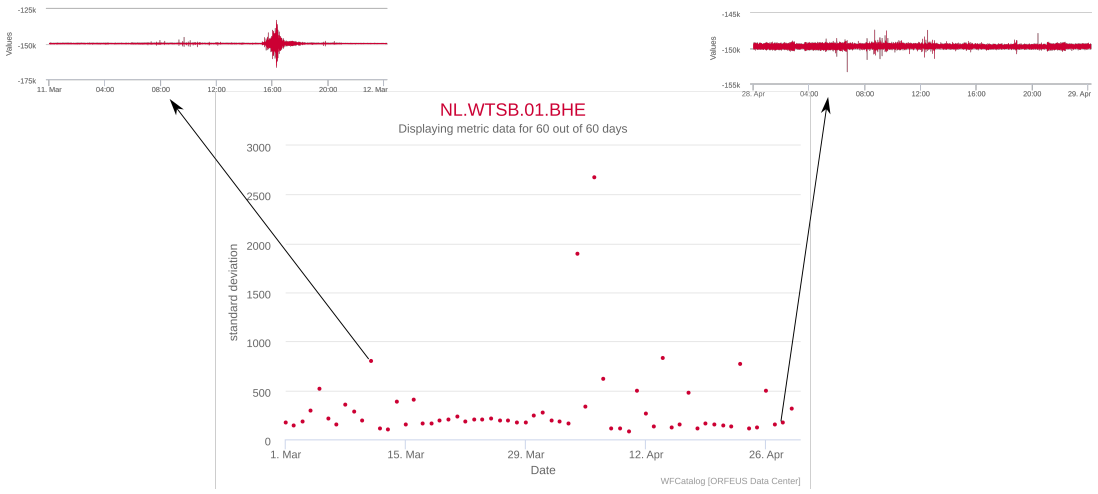
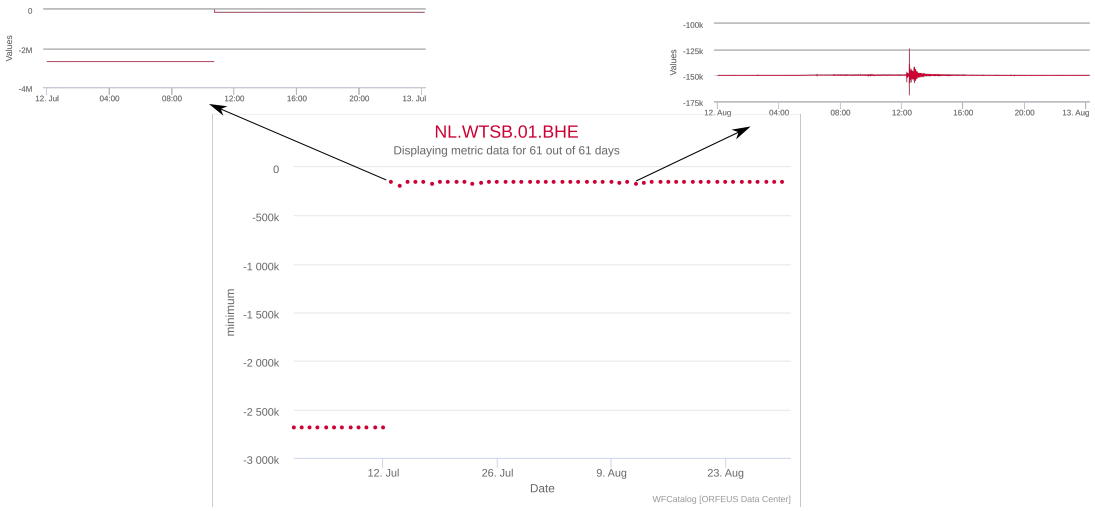


Figure 4.2: Data availability visualisation – this graphical interface allows users to browse through a daily calendar and view data availability. Green and red colours represent high and low availability, respectively. A tile marked with a star indicates full channel availability. Non-continuous days can be clicked to investigate the available data segments for that day[18].

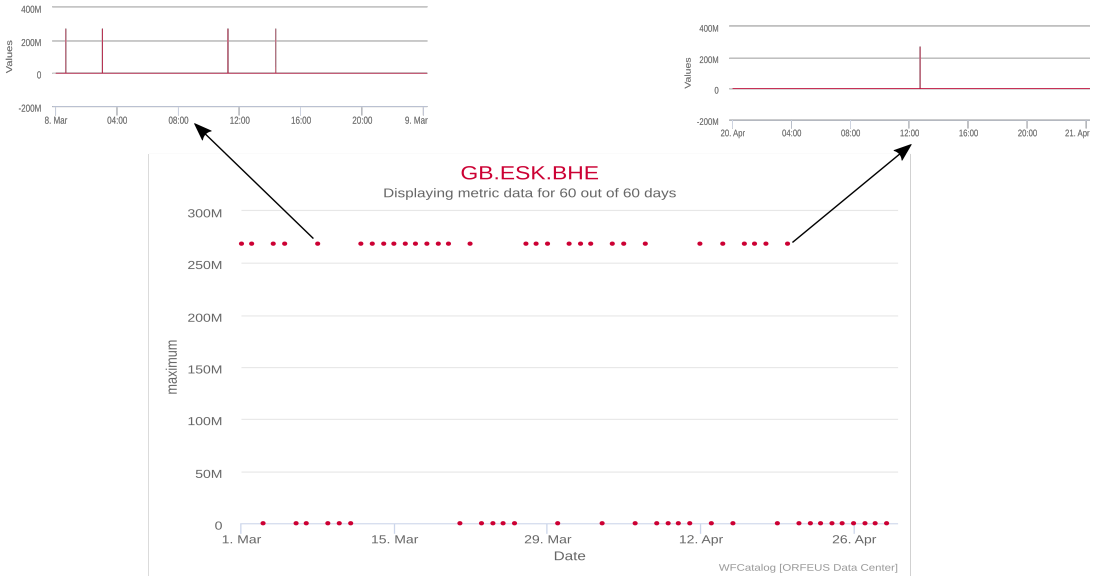


(a) Showing the standard deviation of the raw data. A standard deviation higher than average may indicate the detection of an event during that day (left). Lower standard deviations are representative of ambient noise (right).

[18]Source: `www.orfeus-eu.org/data/odc/quality/availability`



(b) Showing the minimum value of all samples for each daily granule. A sudden jump (left) in the minimum, maximum or mean may indicate an offset in the waveform baseline. A small dip in the minimum (right) may be a feature introduced by an event.



(c) Showing the maximum sample value of the daily granules. Abnormally high maximum or minimum values are indicative of spikes in the data (left and right).

Figure 4.3: Data metric visualisation – this graphical interface illustrates a collection of sample metrics for each day in the requested time window.

4.5.1 Evaluation

We evaluated different aspects of `WFCatalog` discussed below.

4.5.1.1 Metadata store statistics

At present (November 2016) the `WFCatalog` at ODC has information on roughly 4 million daily streams accounting for a total of 400 million continuous segments. The storage size of the metadata of the daily streams, including indexed fields is 1.22 GB using the WiredTiger storage engine available in MongoDB. The metadata about continuous segments accounts for the most significant usage of disk space with a total of 85 GB. The amount of data that is made selectively accessible via this metadata is approximately 15 TB distributed through 4 million daily waveform files. The storage size comes down to a compressed 315 bytes for each daily stream and 83 bytes for each additional continuous segment. Poor waveform data includes many gaps and as a result may consist of up to 500,000 individual traces and will be a strain on the database. Future limits on the minimum length for a continuous segment may be set to prevent explosive growth of the database. In Chapter 6 we provide updates about these statistics. Here we anticipate that a significant growth of the data size ($\tilde{5}$0 TB in October 2018) corresponds to a proportional growth of the metadata size where the size of the continuous segments remains predominant.

4.5.1.2 Benefits for geoscientists

We investigated the advantages provided by `WFCatalog` for improving data discovery. We conducted a simulation to estimate potential saving prior to deploying `WFCatalog` fully. We analysed a sample of real queries (ca. 400,000) submitted by users to `fdsnws-dataselect`, which is currently the most used service from which to retrieve seismic waveform data. Users can submit time-constrained queries attempting to get the desired data streams. However, data delivery is not guaranteed because no *a priori* information about data availability is provided by this service.

`WFCatalog` can be used to get the availability information. We show that by exploiting `WFCatalog` we are able to improve data retrieval and reduce the number of requests which would deliver unusable data. For 'usable' data we mean requested time windows without gaps. By consulting several users we established that this is a likely situation. We submitted the same users' queries to `WFCatalog` with the option to include gaps information in the requested time window.

Figure 4.4 shows the distribution of the requested time windows which have been

analysed. We notice that the majority of the requests are clustered into three main groups: *small range*, *medium range*, *large range*. These groups represent the most popular use cases, which we addressed in our analysis. We compared the responses of the queries with the expected criteria (continuous data), results are shown in Figure 4.5. The percentages indicate the relative gain: $\frac{requests\_with\_gaps}{total\_number\_of\_requests} * 100$. In general there is a substantial improvement in the delivery of correct results as `WFCatalog` informs the users in advance about the time windows that should be discarded without attempting to download them. As expected the benefits increase on larger time windows because there the probability to have a gap is higher.

Therefore by interacting with `WFCatalog` before posing the actual data request, users save time and resources avoiding unnecessary downloads of discontinuous data streams.

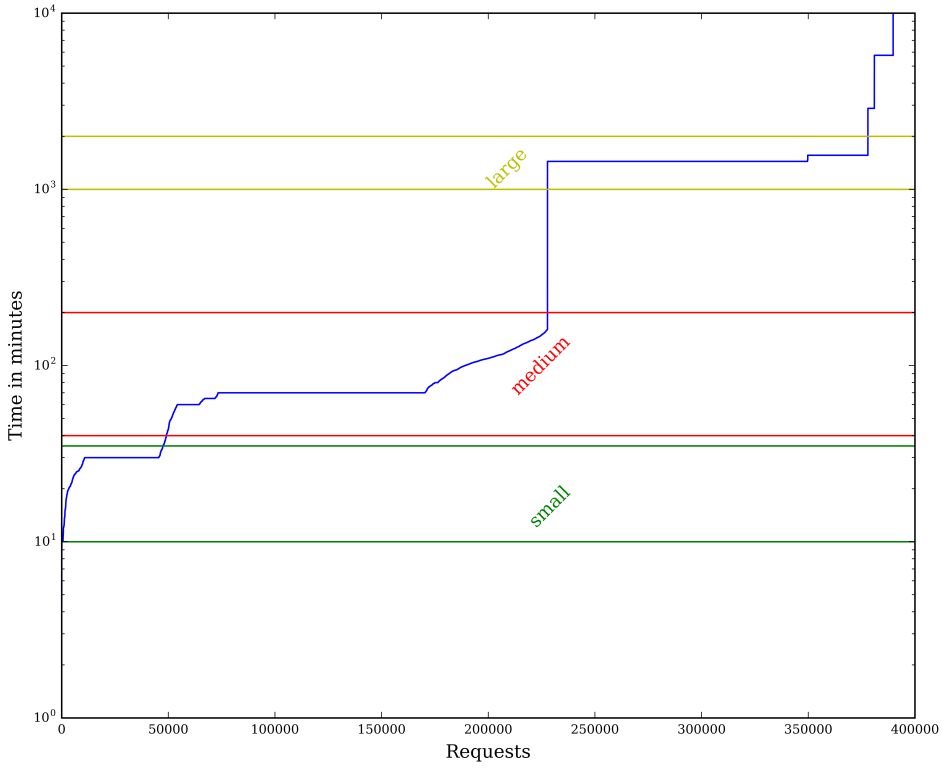


Figure 4.4: Requested time window distribution – the figure represents a sample of real users' requests submitted to `fdsnws-dataselect`. The highlighted regions show the most popular time window lengths.

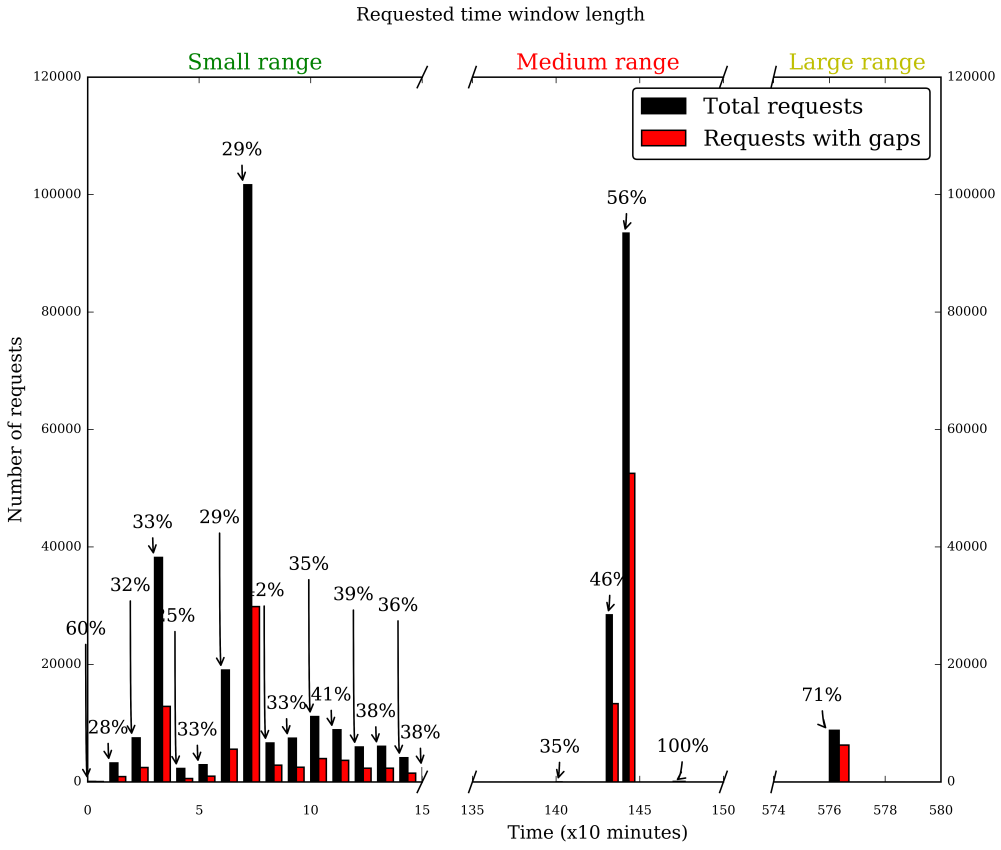


Figure 4.5: Improvement in data delivery: avoiding 'gappy' data – the figure shows the potential gain that can be obtained with `WFCatalog` filtering out time windows with gaps. The percentages indicate the relative gain: $\frac{requests\_with\_gaps}{total\_number\_of\_requests} * 100$

4.6 Conclusions and lessons learned

We presented a novel service which is deployed across the major European seismological data centres of EIDA. `WFCatalog` enriches the portfolio of tools and services for seismology providing clear advantages in the discovery and access of seismic waveform data. The information provided in a machine-readable way will foster automated workflows and improve the data acquisition process. `WFCatalog` with its WFMetadata schema set the basis for a standardised way to exchange seismic waveform metadata and for a canonical representation of quality metrics and data features. The current schema will be maintained and supported by a large community in ORFEUS – this can ensure long term sustainability. Moreover, the continuous interaction with the users

will guarantee extensions in order to address new use cases and scenarios. One such extension is for instance the integration of Power Spectral Density functions which was planned and it has been developed (January 2018) – a tool for the discovery and visualisation of Probabilistic Power Spectral Densities is available[19]. The interoperability with broader communities beyond seismology is another aspect which will be improved, by enriching the published metadata including persistent identifiers and Dublin Core[20]. `WFCatalog` is a step towards making seismic waveform datasets FAIR – 'Findable', 'Accessible', 'Interoperable' and 'Reusable' – [Wilkinson et al., 2016]. We return to consider this effect further in Chapter 6.

The experience acquired in a 'tractable' context described in this chapter was substantial and essential to sharpen our understanding of IPC. It helped us identifying critical challenges. The lessons learned engaging the seismology community shaped our thinking, influenced the definition of our research goals and supported the design of a strategy to tackle them. They corroborated our motivations and encouraged us to continue our investigations to meet the primary goal of this thesis. We can summarise the lessons learned as follows:

- Establishing a focus for collaboration is a costly process which requires active engagement of the stakeholders. Even in a well-organised context the community and the actors involved ought to agree the concepts which need sharing.

- Major issues are not technical but socio-political. Initially we thought that engineering aspects would be predominant. We now recognise that the process of reaching agreement is very demanding and therefore we are motivated to improve it.

- The demand for multi-faceted shared information is growing rapidly driven by: more data, more data sources, the lower cost and complexity of challenges communities tackle. Therefore experts need good strategies to meet the growth. This suggests that we should consider a methodology rather than a one-off solution.

- Such a methodology has to engage experts from user communities and empower

[19] `www.orfeus-eu.org/data/odc/quality/ppsd/`

[20] `http://dublincore.org`

them to directly steer the process with other expert help in the background. Several, equally important aspects are involved – they range from high-level conceptual views to implementation details. Each concern requires in-depth knowledge that often resides with different stakeholders. Loss of interest and disengagement might be triggered by lack of a common vocabulary and diverging interests.

- The collaborative culture ought to be sustained once established as it can help communities respond to new challenges and opportunities more rapidly and effectively. There is a need to support an ongoing process that reviews agreements about information sharing – this requires adequate governance.

Drawing on these considerations we started formulating our strategy – partitioning the challenges into *independent dimensions* can be the key to sustain experts' engagement. We build on a separation of concerns in order to offer usable tools matching the needs of the different stakeholders in order to incentivise and stimulate their interest. We identified three dimensions to be addressed independently: *Conceptual definition* (C), *Representation* (R) and *Population* (P).

In Chapter 5 we discuss these dimensions in detail and leverage them to devise a conceptual framework and a methodology to establish information spaces underpinning IPC.

Chapter 5

Establishing Core Concepts for Information-Powered Collaborations

The content of this chapter is extrapolated from the published article 'Establishing Core Concepts for Information-Powered Collaborations' [Trani et al., 2018a][1]*. Adaptations were made to fit the context of this thesis. Data and statistics are reported unaltered and reflect the status at the time of publication. Related updates are provided in the next chapter as part of the evaluation.*

This chapter presents our approach to building and sustaining Common Information Spaces (CIS) underpinning IPC and its application in a challenging context of a large-scale Research Infrastructure – the European Plate Observing System (EPOS).

In previous chapters we introduced a way to partition our problem space and harnessed such a separation of concerns to review relevant literature. We defined a notation constituted by three dimensions that helped us explore, appreciate and discuss challenges and their implications in each area of concern. We recognised the great value of establishing agreed holistic views over heterogeneous data sources and how they can be enabled by underpinning CIS.

We now describe the dimensions introduced in Chapter 1 in detail and contextualise

[1]The research described in this article has been conceived and designed by myself. I developed the approach, performed the analysis and wrote the paper. Co-authors helped by discussing ideas, by providing inputs and by supporting with the data collection and the data model (EPOS-DCAT-AP) refinements.

them into a conceptual framework that combines two ingredients – collaborations and data – to help establish *Core Concepts* underpinning IPC. We motivate how such a framework can be exploited to lower the barriers for effective cross-disciplinary collaboration and to offer concrete tools to support IPC.

The rationale and key elements of our approach are presented in the next section.

5.1 Building the holistic view

In the introduction of this thesis we motivated the importance of research collaborations and their fundamental role in the advancement and application of science. We introduced the concept of Information-Powered Collaborations (IPC) in order to capture and abstract the complexities and the diverse aspects involved. Cooperation among diverse actors carries inherent socio-technical issues and requires us to maintain "*a common terminology and shared knowledge base*" that enables communication and understanding [Lubich, 1995a]. Therefore, promoting and establishing effective collaborations is a major driver and a key objective of this research and here we outline our methodology to achieve it – this was presented in Figure 1.3 and is reprised with more detail later in this chapter (Fig. 5.2).

From our experiences and assessments (*e.g.* as reported in Chapter 4) it appears clear that the construction of the conceptual framework that enables effective collaboration has to be led by humans. Scientific communities, users and stakeholders of an IPC assume a central role in guiding the construction and maintenance processes. Those shaping the IPC develop and maintain its conceptual core by assessing which concepts can be consistently used and interpreted across the consortium. They often proceed by importing large established vocabularies with their corresponding definitions and relationships – in Chapter 3 we showed how to set up and share such vocabularies. They need to manage the relationships between such *conceptual bundles* eventually extending or pruning them in order to meet the requirements of their IPC. They must recognise where creative diversity exists and leave opportunity for agile innovation in these conceptual spaces.

Our approach combines top-down and bottom-up strategies, or in other words '*meet in the middle*' [Zeginis et al., 2014], to formulate the agreed core set of shared concepts and achieve *semantic interoperability* in IPC. We propose that this progresses

by building a *Canonical Core* (CC) that includes sufficient *Core Concepts* that are agreed and adopted to enable the principal interdisciplinary collaborations to proceed. The extensions needed beyond this CC to support innovation, experiment and local specialisations are supported by dependable relationships with the CC. Approaches based on reference ontologies have been profitably applied in more controlled contexts *e.g.* in the industry [Chungoora et al., 2013; Szejka and Junior, 2017; Imran and Young, 2016]. We build on those results to devise a solution for the challenging IPC context. The whole process exploits *co-design* bringing together data and metadata-modelling experts with domain scientists. Similarly to data models and their representations, the rules of engagement or 'contracts' to participate in the IPC are critical. Such rules are discussed and defined with the designated communities and leverage existing community standards and practices.

Figure 5.1 illustrates an overview of the proposed framework where the Canonical Core is a central component. As stated by the European Commission in its communication on Open Data of December 2011 [European Commission, 2011]: "*[...]the availability of the information in a machine-readable format as well as a thin layer of commonly agreed metadata could facilitate data cross-reference and interoperability and therefore considerably enhance its value for reuse*". In our proposed framework, the Canonical Core captures agreed concepts as machine-readable metadata. The size of the core should account for several factors. It must span a sufficient range of concepts and viewpoints to meet the understood requirements for composing data, information, knowledge and methods. It must offer hooks whereby its capacity may be extended on a local experimental or specialisation basis and recipes or paths to easily incorporate successful extensions. Likewise, the core requires parsimony and consistency to make it comprehensible and manageable.

In the following sections we present diverse aspects or dimensions of the CC.

5.1.1 Dimensions of the Canonical Core

The CC represents the Universe of Discourse that designated communities adopt to communicate, understand and enable 'actionable' information sharing. Actionable in the sense that it can deliver knowledge which can be understood and trusted by practi-

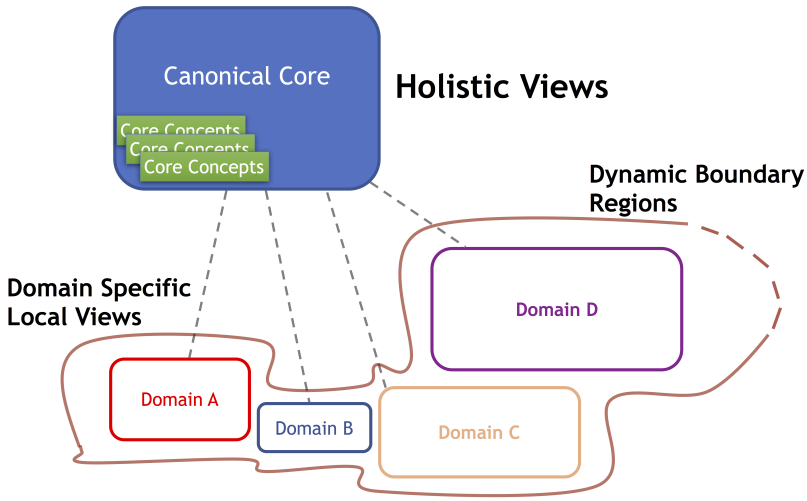


Figure 5.1: Overview of the framework facilitating the composition of diverse resources and their presentation to users as a coherent holistic environment. The CC provides a stable, agreed and adopted set of concepts and their relationships. The need to support innovation and handle details that are not completely adopted is met by recognising external zones of defined information models. Dynamic Boundary Regions delimit the CC from the community-specific extensions.

tioners and interpreted by formalised automated methods. The CC is characterised by three dimensions:

1. *Conceptual definition* – what goes inside the Common Information Space, including the Core Concepts and their relationships.

2. *Representation* – how those concepts are represented, for instance according to a specified data model, *e.g.* using DCAT.

3. *Population* – how the CC is constructed, ingested and maintained with selected instances of the concepts therein represented and how instances are chosen among the available ones (it specifies selection criteria as not everything needs to be included at the highest level of detail).

The conceptual definition (1) constitutes an unbounded conceptual space independent from the other dimensions. In this chapter we provide an approach to manage the complexity of that space, and apply such an approach to the concrete context of

EPOS. We also propose a representation (2) fitting the designed space and meeting the requirements of the identified designated communities. Finally, we validate the chosen representation populated with a selection of real instances (3).

Figure 5.2 illustrates how we combine the three dimensions into a methodology to build and sustain a Canonical Core.

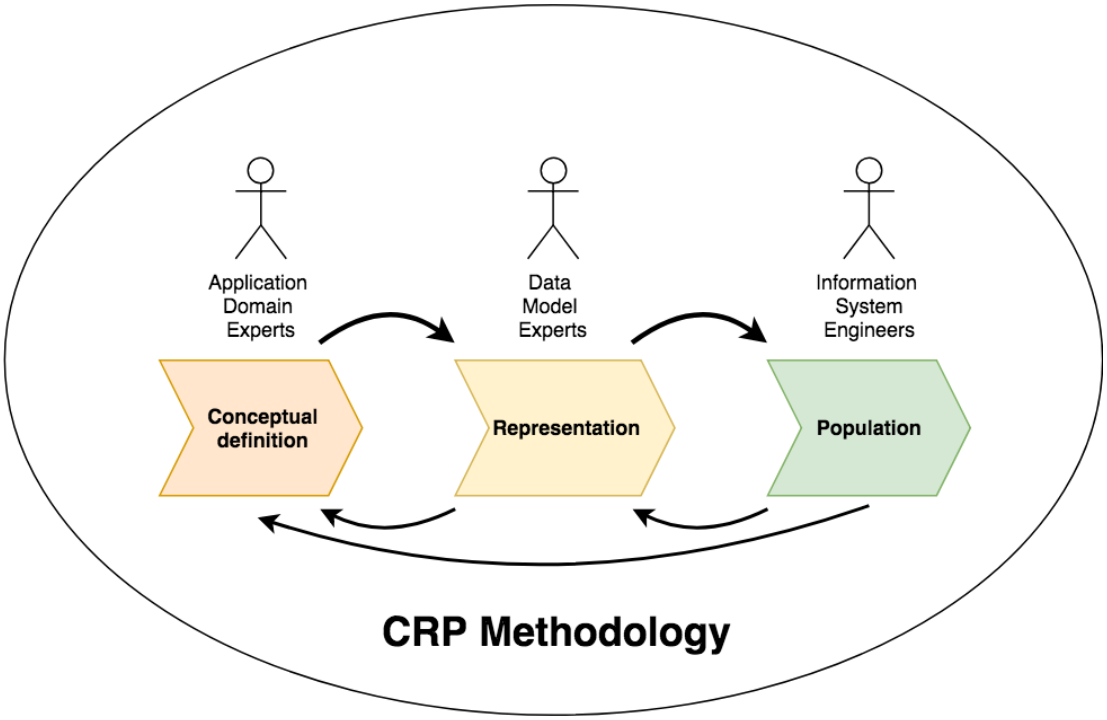


Figure 5.2: Showing the CRP methodology that enables the construction and maintenance of a Canonical Core. CRP is based on a separation of concerns that allows stakeholders to focus on their primary interests. Application domain experts (e.g. scientists, managers) discuss and choose Core Concepts and their definitions. Data model experts co-design appropriate representations for the chosen concepts. Information system engineers are in charge of setting up the population management. The black arrows indicate phase transitions. When conceptual agreements are met requirements are passed to the next phase, *i.e.* representation. Similarly, once the model for representation is defined the population phase can be initiated. Refinements might be required, these are expressed with the backward arrows.

In the next sections we describe the principles underpinning each dimension (*i.e.* C, R and P).

5.1.2 Principles underlying the conceptual definition of the Canonical Core

The conceptual definition of the CC needs to address three aims italicised below. The following principles shape the concepts, relationships and structure of the CC.

1. Achieve sufficient coverage of the behaviours required across the designated communities that the CC supports their interactions with the shared information and with each other, thereby facilitating collaboration leading to *adoption and reuse*.

2. Establish agreed interpretations of the Core Concepts that are adopted by the designated communities – when such agreements cannot be reached allocate the concepts to an extension for the relevant subcommunity coupled to the core via identified conceptual hooks – thereby achieving *harmonisation* without inhibiting innovation.

3. Validate the CC against a broad and representative set of use cases, thereby ensuring priority collaborative behaviours are enabled and achieving *trustworthiness* and *completeness*.

The volume and complexity is controlled by limiting the core to accepted and agreed material. Contenders for inclusion develop in the dynamically connected Boundary Regions. The set of use cases is extended to fulfill all critical requirements and to ensure that the CC covers the essentials.

According to the principle (1), rather than building from scratch we select and import existing conceptual bundles, information spaces, boundary objects and knowledge artifacts [Star and Griesemer, 1989; Star, 2010; Cabitza et al., 2008, 2013] into the CC. This adoption of existing bundles has two motivations: a) to retain intellectual effort – as bundles are often the result of long and costly negotiation (implicit and explicit); and b) to facilitate understanding and automated interaction – as communities and their automated methods will recognise familiar patterns and artifacts.

Nonetheless, the CC cannot be just the union of pre-existing bundles – *harmonisation* (2) plays an essential role. Without harmonisation the CC would be a collection of information silos that preserve domain specific structures together with their boundaries. This would result in a data warehouse that collects data unchanged, thus failing

our principal goal that is to facilitate comprehensible boundary crossing by providing holistic semantic integration.

We harness real *use cases* (3) to tease out and clarify the objectives and aims of the designated communities whose work and communication will be mediated via the CC when they adventure across previous boundaries. To turn an unbounded conceptual space into a manageable space we follow communities' priorities. As use cases evolve and change the associated dependencies and boundaries follow accordingly, thereby identifying required extensions and modifications to the core. Hence, the CC has a clear requirement for flexibility and support for evolution. These guiding principles shape the construction and evolution of the Core Concepts.

5.1.3 Principles underlying the representation of the Canonical Core

Representation entails metadata, as we discussed in Chapter 3. It reflects aspects of the real world for intended purposes and viewpoints [Alemu and Stevens, 2015; Gartner, 2016]. The representation of the CC requires appropriate metadata to describe the complexity of IPC for their supported use cases. As the CC needs to accommodate heterogeneous bundles typically with different encodings, the representation of the core must support what Nilsson called *horizontal harmonisation* [Nilsson, 2010], that is interoperability across different standards. We adopt the principles for enabling interoperability defined by Duval et al. [Duval et al., 2002], recalled in the Memorandum of Understanding between the Dublin Core Metadata Initiative (DCMI) and the IEEE Learning Technology Standards Committee (IEEE LTSC) [DCMI, 2000] and extended by Nilsson et al. [Nilsson and Johnston, 2006; Nilsson, 2010]. These deliver the following:

1. *Extensibility*, ability to create and add new structures to a metadata standard for "application-specific or community-specific needs".

2. *Modularity*, "ability to combine metadata fragments adhering to different standards".

3. *Refinements*, "ability to create semantic extensions".

4. *Multilingualism*, "ability to express, process and display metadata in a number of linguistic and cultural circumstances".

5. *Machine-processability*, "ability to automate processing of different aspects of the metadata specifications".

These principles fit the characteristics of an IPC as they assume and acknowledge the co-existence of multiple standards and different specifications. Also, they enable the collaborative approach, for instance members of the designated communities can annotate existing content creating new relationships (1) and refinements (3). Moreover we identify additional issues to consider:

6. *Minimal ontological commitment* is sufficient coverage "to support the intended knowledge sharing activities" without introducing unnecessary ontological terms [Gruber, 1995]. That requires us to specify "only those terms that are essential for the communication" of consistent and understandable knowledge. It advocates underspecification, specialised meanings being introduced via extensions.

7. *Maturity and level of standardisation* provide a measure of the acceptance among communities as well as an indication of the investments made for uptake. In particular they are reflected in (a) the number of bundles already encoded in a specific representation; (b) the set of available tools compatible with such a representation; and (c) the support offered by communities of experts.

8. *Expressivity and richness* as the ability and the easiness to express logical relationships are important factors that influence the choice of the representation for specific use cases.

9. *Effectiveness* representing the required concepts for the selected application scenario. For instance, verbosity might be more effective in machine-to-machine exchanges whereas terseness might help human reading and understanding. *E.g.* RDF/XML [Gandon and Schreiber, 2014] is an example of verbose representation of RDF whereas Turtle/RDF [Prud'hommeaux and Carothers, 2014] and N3 [Berners-Lee and Connolly, 2011] are terse and readable representations.

10. *Performance* of the encoding/decoding processes, required to marshall and unmarshall the content of the core. This is an important non-functional engineering

aspect that influences the overall behaviour of the system and in particular of the population described in the next section.

11. Support for *validation and consistency* checks. This can be achieved by adopting formal restrictions, constraints, description logic, formal rules and inference mechanisms *e.g.* XML Schema, OWL, SHACL and SPARQL-based validation.

5.1.4 Principles underlying the population of the Canonical Core

The population describes the distribution in time of the entities (instances of concepts and instances of relationships between them) in the CC. Population is a dynamic process that is guided by the principles listed below.

1. The *strategy* adopted to populate the CC is influenced by several factors *e.g.* volume of data, restrictions, governance. However, the possible approaches are: (a) reference or *brokering* – pointers to externally managed bundles are stored in the CC; (b) copy or *harvesting* – the CC holds a physical copy of bundles; and (c) mixed – a combination of the previous two where the CC holds a physical copy of a subset of a bundle, *e.g.* of the information used first or most frequently. As reported in Chapter 3 the OAIS RM contemplates similar strategies which are referred to as *referencing* and *collecting* respectively.

2. Related to the population strategy are the concepts of *conceptual* or *logical* population and *actual* population. The logical population indicates the number of entities which are potentially made available by the CC, whereas the actual population indicates the number of entities currently available. This observation introduces the concept of *latency*, which is the time required to move from the logical to the actual population. For instance, the CC might contain pointers or references to entities of an external catalogue. Although these external entities logically belong to the CC, and thus they are available for the users of the IPC core, there might be a delay to provide access to the concrete objects represented by those entities. In Chapter 3 we presented ways to keep distributed populations synchronised.

3. *Quality control* is fundamental to manage the population of the core. Quality indicators must be used to assess new entities and providers of entities as well as

to modify the population, for instance by removing entities that do not conform to defined quality standards. Pruning, clean-up, deduplication and notification mechanisms should be implemented exploiting such quality indicators.

4. *Governance* – community endorsed decisions on the target populations and their management. For instance, existing community agreements associated with specific bundles might influence the population strategy and require access control mechanisms.

5.1.5 A note on governance

We mentioned governance as one of the aspects influencing the population dimension. For the sake of clarity, we should notice that governance plays a major cross-cutting role in all the components of the framework. For instance, in the conceptual definition governance might influence the strategies to manage the size and content of the CC and adjudicate on what is ready for inclusion in the next release of the CC. It would define priorities when addressing users' requirements and choosing relevant use cases. In the representation, governance would influence the choice of supported data models and related formats. In the population it would be involved in defining and maintaining agreements between participants of an IPC in order to guarantee arranged quality of service *e.g.* via SLAs. Providing a thorough analysis of governance aspects would require considerable investigations that are out of the scope of this research. We limit our focus by observing beneficial effects of organisational frameworks underpinning IPC as illustrated in Chapters 2 and 4. Although it is not our goal to define specific governance models, we acknowledge that coordinated commitments and sharing of responsibilities are required. By partitioning the challenges our framework allows stakeholders to focus on specific aspects with clear boundaries thus lowering the barriers to exchange and mutual understanding. It offers a concrete tool to foster good and disciplined behaviour thereby achieving an effective collaborative culture. It promotes shared responsibilities in each dimension that nonetheless ought to be arranged, resourced and sustained.

5.1.6 Considerations about the boundary regions

In the previous sections we focused our analysis on the characteristics of the CC, we briefly mentioned Boundary Regions (BR). The CC is an abstraction layer avoiding the complexity of the BR – the core falls under a federation-wide governance whereas BR are independently controlled. For this reason it is difficult to provide a full characterisation of BR. Therefore, our focus is at the *interface* between the boundary regions and the core and on the 'rules of engagement'. Such rules can be modelled leveraging the 'boundary objects' concept introduced by Star and Griesemer [Star and Griesemer, 1989; Star, 2010] and reported in Chapter 2. The authors propose a mechanism to represent and exchange knowledge across organisational borders that should facilitate communication. We have seen further CSCW literature built on that concept. Cabitza et al. introduce the concept of 'knowledge artifact' that provides *bounded openness*. It "*allows participants to establish a shared meaning on the one hand, while remaining open for modifications on the other*" [Ackerman et al., 2013; Cabitza et al., 2013].

Below we list characteristics of BR that provide the requirements for the interface with the core.

1. BR generate both requirements and constraints for the CC. Such requirements and constraints are time dependent and have a high variation due to the inherent dynamic nature of the regions. Hence, the interface with the core ought to accommodate such *variations*.

2. BR expose a bounded-openness – new boundary regions can be added, removed and at the same time each region can contribute new bundles to the core, provided they fulfil the agreements negotiated with the core.

3. Popular bundles are easily recognised, connected and imported into the core, as they typically gather consensus and form standards whereas less popular bundles constitute extensions. The value of both must be preserved and accounted for, thus the interface has to support both cases and allow differences. In 1945 Vannevar Bush describing memex, wrote "*trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory*" [Bush and Wang, 1945]. This captures very well the requirement for promoting and

highlighting *extensions* based on diverse criteria in order to engage and attract users and avoid unproductive migrations to other systems, dispersions and so-called 'skunk work', where researchers hide their activities to achieve agility and flexibility with consequent loss of evidence for reproducibility and sharing.

To address these requirements the interface between the core and the boundary regions can be modelled as an API for managing extensions. Such an API supports the following operations: 1. registering an extension and holding information about creator and responsible party; 2. noting the aspects of the CC on which an extension depends; 3. winding up work on an extension; and 4. adopting (parts of) an extension into the core.

The following example illustrates how such an API would work in practice. A subcommunity (*SubCom*) harnesses a subset of the CC (C_{sub}) to conduct experimental investigations that yield new data and related concepts, a new conceptual bundle (C_{new}). C_{new} gains respect and interest from other research groups who would like to use it as early adopters. In order to make it accessible, the API registers C_{new} collecting information about *SubCom* and C_{sub}. When *SubCom* has completed the experiments a new (stable) version of C_{new} is available, $C_{new.stable}$. Depending on the relevance or other criteria $C_{new.stable}$ (or parts of it) might be promoted as new bundle in the core. This scenario has implications on the core and calls for additional requirements such as: *versioning* and *provenance*.

5.2 Applying the CRP principles – a practical approach

In the previous sections we introduced the three aspects of the CRP methodology and described their underpinning principles. Here we outline our approach to translate those principles into practice thereby defining an initial set of the CRP's processes. The application of such processes in a specific context is described in the next section (5.3). We present the activities that can be performed to fulfil the CRP principles organised by the three phases of the methodology (*i.e.* C, R and P) and discuss possible improvements.

5.2.1 Conceptual definition

To enable the definition of Core Concepts adhering to the principles illustrated in Section 5.1.2 we propose the following activities.

1. Identify and engage key communities' roles or '*community gateways*'. This can be achieved by organising meetings and by establishing direct contacts (*e.g.* by email). It is a time and resource consuming activity which requires participation and commitment of the identified parties. Preparation work might be required, *e.g.* to prepare informative material and present introductory talks, to illustrate the approach and to plan allocation of resources.

2. Collect requirements and use cases. This activity can be performed in stages. For instance, preliminary templates can precede interviews and detailed questionnaires. Accuracy and completeness of responses might be quite variable and require further elaboration and interactions.

3. Survey existing information the IPC needs to share including information about data, practices, methods, resources, individuals, organisations, providers, contact points and descriptions. A community gateways' knowledge can be harnessed to pinpoint additional and more specific information. Shared documents, spreadsheets and templates are useful tools in this phase.

4. Prioritise use cases and assets. This activity requires strong engagement with the domain experts and final users of the Canonical Core who should decide and set their own priorities.

5. Derive an initial set of Core Concepts *e.g.* by analysing collected material. This phase can be conducted by knowledge engineering experts who can then ask for domain experts' feedback.

6. Perform a preliminary classification of the concepts *e.g.* by creating possible categories. This depends on domain experts with the support of knowledge experts. Visual modelling tools such as mind maps are useful at this stage.

7. Initiate harmonisation processes. This phase requires direct interactions and exchanges between participants of the IPC. Dedicated workshops supported

by knowledge experts are useful to organise focused task forces and working groups. Harmonisation is a time consuming activity that should be addressed by successive refinements. These can be conducted in stages alternating offline work and meetings.

8. Define representative and agreed definitions. These are the results of harmonisations and experts' interactions. In this phase bundles already used by parts of the IPC could be imported and adopted. Shared documents might be harnessed to capture and discuss definitions. Similarly, established vocabularies might be consulted as valuable sources for existing terms.

9. Refine classifications and relationships. The information acquired and analysed can be used to refine the organisation of Core Concepts and their relationships. Modelling tools such as conceptual class diagrams help in this phase.

It is worth noticing that the presented activities could be performed in a different order or advance in parallel and in some cases they might overlap. Also, iterations are usually needed to achieve a first representative set of Core Concepts. Possible improvements include the application of established practices for requirements collection and classification and ontology engineering methods such as competency questions and scenarios [Suárez-Figueroa et al., 2015].

5.2.2 Representation

The representation principles can be fulfilled by performing the following activities.

1. Collect ontological and non-ontological resources. They offer a more precise view of the assets by providing their possible representations and formalisations. Shared documents set up during the conceptual definition phase can be enriched with this additional information.

2. Analyse existing vocabularies, standards and popular models. This phase is important to foster re-use and avoid duplication. Search engines and tools such as the Linked Open Vocabularies (LOV[2]) can be adopted to look up terms and their definitions.

[2] https://lov.linkeddata.es/dataset/lov/

3. Assess applicability of existing bundles in the intended context. In this phase selected concepts are tested against the identified use cases and definitions. When they fulfil the requirements they can be imported as existing bundles, otherwise customisations and redefinitions are needed. In the latter case linking extensions to related concepts in existing vocabularies could be beneficial to facilitate uptake and understanding.

4. Design the model including details such as entities, properties and relationships. This can be performed by building around clusters of previously selected bundles and by performing extensions when needed. Data-modelling experts lead this activity and can make use of notations such as UML class diagrams.

5. Validate the model against requirements. Rather than performing a formal validation, the goal of this phase is to check that the principal use cases are covered by the defined entities and relationships. This requires consultation with domain experts.

6. Encode the validated model in one or more formats including structures and constraints. The choice of the level of specification is crucial. It is important to balance consistency and flexibility. For instance, too strict constraints might affect application and reuse of the model. This phase should be guided by the minimal ontological commitment principle. Advanced editors and platforms such as Protégé [Musen, 2015] can provide significant help.

7. Test the encoded model with sample data. This can be achieved for instance, by means of prototype implementations and/or mock-ups with users.

8. Establish a review process. A systematic approach to address issue management is required.

9. Publish and explain the representation to representatives of the IPC, users and other stakeholders and gather feedback and comments. In this phase the representation is shared with data model experts and domain experts. Collaborative tools such as GitHub or GitLab will help tracking changes and issues.

10. Prepare and organise a release when the representation of the Core Concepts

is sufficiently detailed to address requirements and expectations. The release process should include preparing documentation and training.

The activities presented in this section are mainly driven by data modellers and information experts. Domain experts are consulted to address specific issues and to be involved in the review processes. Structured approaches and tools (*e.g.* ontology engineering) could offer significant support and improvements [Ontology Engineering Group, 2019; Alobaid et al., 2018].

5.2.3 Population

The following activities enable the realisation of the principles underpinning the population dimension.

1. Identify and collect data and information sources. Each concept represented and shared in the IPC should correspond to one or more sources of information. These provide the content and values to be associated with Core Concepts. This phase requires interaction with communities' engineering experts who are consulted to exchange knowledge about service endpoints, API, protocols, media types, formats, *etc.*

2. Specify selection criteria for the instances of the concepts. After the sources of information to create Core Concepts instances are identified selection rules are needed *e.g.* to filter out unnecessary content. In this phase such rules are defined, formalised (*e.g.* via import-transformations, mappings and conversions) and associated with the corresponding sources. Shared documents can be used to maintain this information.

3. Develop templates and examples of usage. These help initiate the population process and familiarise users with the representations developed in the previous phase. In its initial stages population is usually a manual, human-intensive activity that can be automated after practices are assimilated and tested. Collaborative environments can be adopted to support interactions between communities' representatives and data-model experts.

4. Organise dissemination activities such as workshops and training. Hands-on sessions involving communities' developers and engineers are particularly useful

in this phase. For instance, by working jointly on the mapping of community assets into the chosen representations the type of work and effort required becomes evident, procedures are clarified, issues are exposed and preliminary feedback collected. Webinars covering focused topics can be valid alternatives to face-to-face meetings and can help reach a broader audience.

5. Initiate population (*e.g.* starting with selected entities). Following up on the dissemination activities communities start working on the population of their specific entities. This activity creates concrete instances of Core Concepts and typically requires considerable effort and time. In order to make this task more manageable effective planning could include prioritisation of entities and population cycles with defined targets.

6. Curate inputs and provide feedback. The curation of the values provided by the communities is crucial. It allows those involved in the population to verify the correctness of the information, to assess understanding of the processes and to perform adjustments where needed. Automated tools are useful to check syntactical and structural errors *e.g.* RDF SHACL validators. Visual tools can support content navigation and inspection, nevertheless this is currently a human-driven activity.

7. Define population strategies. Once the population process has been tested and validated, automation can start. However, population strategies (*e.g.* brokering, harvesting) have to be associated with the particular sources of information and the corresponding entities. Those choices should be recorded and captured *e.g.* in shared documents and spreadsheets, ultimately they should be supported in the chosen representation.

8. Design and implement the population architecture. Modular and extensible frameworks facilitate the integration with existing community software, tools and services. Co-design and co-development involving community engineers and data model experts are advisable. For instance, convertors and parsers can be built to extract the required information from existing community standards. Mapping tools and declarative languages such as RML [Dimou et al., 2014] can provide significant support in this phase.

9. Prepare for operation. The technical infrastructure requires organisational support and governance. Agreements with providers should be defined and formalised to sustain the population processes in the operational phase. Those agreements (*e.g.* SLA) should include indicators to assess quality and define roles and responsibilities.

10. Launch operation. This last phase requires that technical and organisational frameworks are established to manage the operational processes. For instance, monitoring is required to make sure that the system behaves according to agreed policies. Modifications and changes of the populations need to be negotiated with the governance of the Canonical Core.

Setting up the population phase involves mainly domain engineers and experts of data models who work jointly in the development of the population architecture. When moving towards operation the role of governance becomes more relevant. Population remains an ongoing activity that enables the evolution of the Canonical Core according to defined policies.

In the next sections we present a concrete example of application of the CRP methodology in EPOS.

5.3 Building the EPOS Canonical Core

In this section we describe an application of the approach introduced in the previous sections. We apply our methodology to establish the EPOS CC addressing its three dimensions: conceptual definition, representation and population. The proposed implementation leverages and integrates several of the technical solutions presented in Chapter 3. We describe it in the next sections after introducing the context of our application scenario.

5.3.1 European Plate Observing System (EPOS)

The European Plate Observing System (EPOS[3]) is building a pan-European research infrastructure for solid-Earth sciences. It will start its operational phase in October

[3] www.epos-ip.org

2019 with the establishment of an European Research Infrastructure Consortium (ERIC). The mission of EPOS is to integrate the diverse and advanced European Research Infrastructures for solid-Earth sciences creating new opportunities to monitor and understand the dynamic and complex solid-Earth system [Bailo et al., 2018].

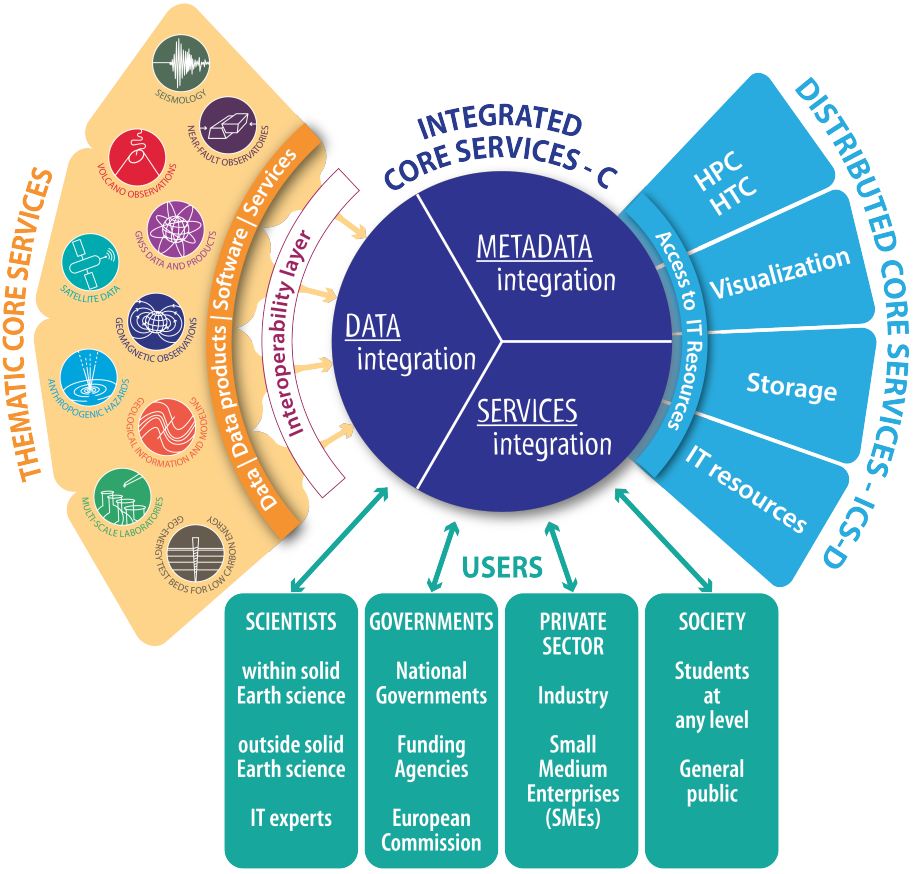


Figure 5.3: European Plate Observing System organisational architecture high-level overview – the EPOS-CC is hosted in the ICS-C. It is represented and maintained in a central metadata catalogue that supports the offered services by steering interactions and information exchanges.

EPOS is a prominent example of an IPC that targets ten different scientific communities: seismology, near-fault observatories, Global Navigation Satellite System (GNSS), volcanology, geomagnetic observations, geology, satellite observations, anthropogenic hazards, multi-scale laboratories and geo-energy test beds. It currently (March 2018) involves 141 institutes and organisations spanning 22 countries and connects with the

global Earth-observation communities. For complexity and scale EPOS provides a rich set of requirements and challenges typical of IPC. Figure 5.3 depicts a conceptual view of the socio-technical architecture supporting EPOS. Such an architecture is composed by three fundamental elements (from left to right):

- *Thematic Core Services* (TCS) – these are provided by the participating communities. Within each of the targeted domains EPOS has promoted and stimulated the harmonisation of data management, access methods and policies, as well as services (*e.g.* processing, visualisation) and resource provisioning by: 1. fostering the creation of new European-wide thematic hubs; and 2. supporting existing organisations (*e.g.* ORFEUS[4] for seismology). However, much intrinsic diversity remains.

- *Integrated Core Services* - Centralised (ICS-C) – they constitute the novel system under construction to integrate the diverse resources provided by the TCS. Interoperation between the ICS and TCS is needed. This requires the description of available resources by means of rich, flexible and standardised metadata. It supports the data life-cycle from acquisition to exploitation and the conduct of scientific methods and sustained research campaigns.

- *Integrated Core Services* - Distributed (ICS-D) – they constitute the distributed part of the ICS. These services are offered by e-Infrastructure providers and resource providers that – under clear procurement policies or SLAs – make resources available (*e.g.* HPC, HTC, data storage and data transport) for the operation of the ICS's computational or visualisation tasks.

ICS-C and ICS-D are grouped logically into one component which we refer to as ICS. The metadata describing data and assets are hosted in the EPOS ICS Metadata Catalogue (EIMC). The EPOS CC is represented in the EIMC that underpins the organisation of integration processes and fosters interoperability between the multidisciplinary data, products, software, services and resources of the contributing research communities.

[4] www.orfeus-eu.org

5.3.2 Definition of the EPOS Canonical Core

The initial definition of the EPOS CC has been conducted by the EPOS metadata group[5] (that includes diverse expertise across the ten themes and informatics) based on a set of requirements and use cases collected during the FP7 EPOS-PP (Preparatory Phase) and H2020 EPOS-IP[6] projects, according to the principles presented in Section 5.1.2. As EPOS is building an infrastructure on top of existing assets, reuse and adaptation is essential. A specific task was the production of a survey of existing resources contributed by the EPOS designated communities. That survey leveraged: 1. the RIDE database[7]; and 2. internal reports from focused campaigns with the EPOS communities.

The survey collected information such as providers, contact points, descriptions of resources, and delivered a preliminary classification of the resources in four categories, namely: Data, Data Products, Services and Software (DDSS). Each community contributed a prioritised list of resources to be included in the core based on their maturity and relevance. For instance, the seismological community provided a set of standardised web services[8] (*e.g.* FDSNWS and EIDAWS), primary data (*e.g.* seismic waveform and strong motion data) and data products (*e.g.* earthquake catalogs and hazard maps). Examples of resources by other communities include: InSAR displacement maps, geochemical data, geological maps, meteorological parameters. Figure 5.4 shows a summary of the current (March 2018) DDSS elements.

The DDSS survey is a valuable asset given the wide scope and heterogeneity of EPOS. A strong engagement strategy with the communities exploiting several communication channels was required in order to agree it. Starting from the DDSS a finer-grained classification has been produced with incremental refinements leading to the definition of the EPOS CC. Such refinements were influenced by geospatial standards (*e.g.* ISO19115) and the CERIF data model [Jeffery and Bailo, 2014; Bailo et al., 2017]. The current EPOS CC includes concepts such as: Dataset, Equipment, Facility, Organisation, Person, Publication, Service, Software, WebService and Project. These are described in Table 5.1.

[5] I was coordinating the activities of this group

[6] `https://epos-ip.org/`

[7] `www.epos-ip.org/ride`

[8] `http://www.orfeus-eu.org/data/eida/webservices/`

Concept name	**Description**
Person	A person: dead, alive, real or fictional.
Organisation	Any type of organisation, *i.e.* social institution. *E.g.* research institution, government agency, corporation.
Dataset	Structured information describing some topic(s) of interest, typically available for access or download in one or more formats.
Equipment	An instrument, a devise used for research purposes. *E.g.* a measuring devise, a sensor.
Facility	A physical or logical place where research is conducted. It includes resources, services, equipments *etc.* *E.g.* a laboratory, a library.
Service	A service provided by an organisation. *E.g.* access service, accounting service.
WebService	A specific type of service programmatically accessible over the Web.
Publication	A scholarly form of creative work *e.g.* scientific, academic publication.
Software	A program instructing a computer to perform specific tasks.
Project	Individual or collaborative enterprise with a planned aim, start date, duration, budget *etc.* *E.g.* a research project.

Table 5.1: EPOS Core Concepts and their descriptions

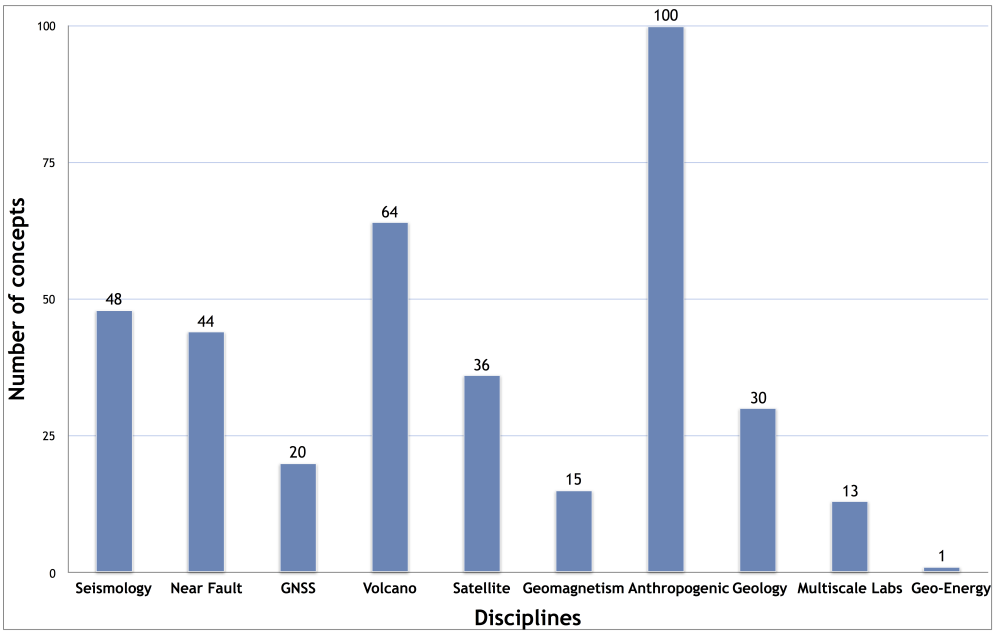
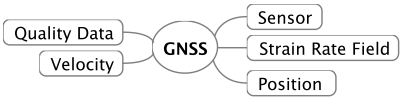


Figure 5.4: Number of concepts of Data, DataProducts, Services and Software (DDSS) offered for sharing by each thematic area

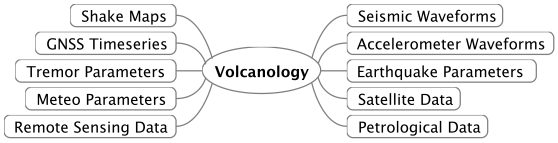
Community bundles (*i.e.* sets of concepts) are made accessible by linking them to the EPOS Core Concepts: *e.g.* $SeismicWaveform \rightarrow Dataset$. Figure 5.5 offers some examples of such bundles. An overview of the existing resources triggered the *harmonisation* process aimed at providing consistent definitions and interpretations across the EPOS designated communities. Commonalities emerged between diverse disciplines. The DDSS survey revealed overlapping areas across disciplines and highlighted variations in interpretations. For instance, the concept of *Seismic Waveform* is shared across a number disciplines besides Seismology *e.g.* Volcano Observations and Near-Fault Observatories. Similarly the notion of *Event* is quite broadly accepted and in common usage among the communities, however, in some cases there is the need to redefine and/or specialise it – for instance the Anthropogenic Hazards community developed and adopted a slightly different and related definition, which they refer to as an *Episode*.

Such examples provide an insight into the typical issues arising in multidisciplinary collaborations. The collaborative work initiated in this process have yielded important

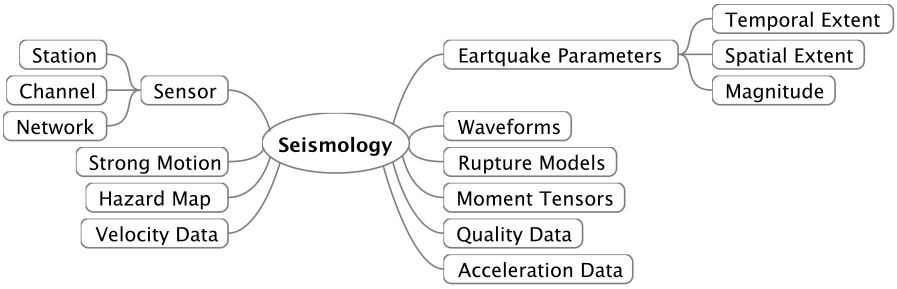
results. It has stimulated and encouraged communities to (re)think about their internal knowledge structure, organisation and formalisation. It has fostered the development of shared controlled vocabularies and taxonomies by forming dedicated task forces with a beneficial exchange of expertise. Communities with traditionally more expertise about classifications and knowledge organisation systems, *e.g.* Geology, shared their approaches with communities less experienced in those topics. Another important outcome was the identification of representative definitions as well as authoritative sources responsible for each specific bundle and set of entities. This allowed EPOS to avoid duplicated definitions and to provide accurate 'reference' definitions with the corresponding representations of the entities. Variations on the reference definitions are allowed where needed but they then need to be linked to and grouped with the reference definitions.



(a) Example of a conceptual bundle from the GNSS community – GNSS concepts can be applied in many contexts *e.g.* to estimate volcano deformations and seismic displacements



(b) Example of a conceptual bundle from the volcano observations community – this community is a predominant example of exploitation of multi-disciplinary, crosscutting concepts



(c) Example of a conceptual bundle from the seismological community

Figure 5.5: Examples of community bundles – Noteworthy is the presence of overlapping concepts whose definitions might be adopted unaltered by a different community (*e.g.* seismic waveform). However, specialisations, modifications and partial reuse have to be accounted for. In some cases similar concepts may have different interpretations (*e.g.* quality data). The CC has to accommodate diversity and support a range of required scenarios.

The concepts and entities collected in the CC underpin the use cases and requirements developed by those supporting the IPC. In Chapter 1 we provided some examples addressing resource discovery, resource evaluation and workflow support. The EPOS CC definition is an ongoing process that will continue after EPOS has transitioned to its operational phase [Cocco, 2018]. The conceptual framework established and described here will be a valuable tool to support the evolution of this core. For instance, it can facilitate the discussion about how to deal with domain specific knowledge and to set up criteria and policies to manage the promotion of community concepts into the CC. We applied this approach in EPOS and defined a process to manage the evolution of the CC – it is presented and evaluated in the next chapter.

5.3.3 EPOS Canonical Core representation

After completing the conceptual definition of the first version of the EPOS CC, the next step was to find a suitable representation that would meet the requirements of the designated communities following the principles in Section 5.1.3. The CC needed to be formalised in this notation to support a) human communication about the concepts of the core, and b) automated processes assembling, managing, accessing and translating entities corresponding to those concepts.

Along with the overview of the communities' assets, information was collected about ontological and non-ontological resources *e.g.* the formats, conventions, vocabularies and standards adopted by the communities to represent their resources. In particular the survey revealed that several domain-specific standards co-exist with broader standards. The adoption of standards and shared practices depends on the maturity of the communities. They can be quite heterogeneous. Table 5.2 provides an example of such diversity. More mature communities follow well-established and broadly applied standards and policies, whereas less mature communities in EPOS needed to initiate standardisation and consolidation procedures. The residual inherent heterogeneity is reflected in the composition of the CC and provides additional constraints when choosing a feasible representation. Noteworthy is the adoption of metadata standards for spatial information such as ISO19115, ISO19139, the OGC standards[9] and the INSPIRE conventions [EU Parliament, 2007] *e.g.* by the Geological modelling community.

[9] http://www.opengeospatial.org/standards

Encoding	Scope	Discipline(s)
MiniSeed	Global	Seismology
WFMetadata-JSON	Community	Seismology
QuakeML	Global	Seismology
Shakemap XML	Community	Seismology
OGC WFS	Global	Geology
OGC WMS	Global	Seismology, Geology
OGC CSW	Global	Geology
CKAN-JSON	Community	Laboratories
Magnetic-HTML	Community	Geo-Magnetic Observatories
OpenSearch XML	Global	Satellite
VpVs-JSON	Community	Near-Fault Observatories
Radon-JSON	Community	Volcanology
CO2-JSON	Community	Volcanology

Table 5.2: Examples of encodings used in EPOS bundles – it provides an overview of the scope and heterogeneity in formats and the adoption of both community and global standards.

A representation has been proposed for the EPOS CC building on the DCAT W3C recommendation, namely the EPOS-DCAT-AP. The DCAT data model is represented in RDF, it supports the principles of Linked Open Data (LOD) and reuses concepts from existing vocabularies. Therefore it meets the principles in 5.1.3. To fulfil the EPOS requirements an EPOS DCAT Application Profile, inspired by Geo-DCAT-AP [European Commission, 2015b], has been developed extending the general DCAT data model. It follows the recommendation on DCAT-AP extensions [PwC EU Services, 2017] and addresses the following concerns:

- Extending the data model with additional concepts required by the EPOS CC; *e.g.* Equipment, Facility, Publication, Service, WebService, Project, Operation, SoftwareApplication and SoftwareCode.
- Introducing new relationships and roles; *e.g.* `epos:resource` that extends the scope of a Catalog in order to a include broader set of catalogued resources (beyond Datasets), `schema:affiliation` and `schema:owner`.
- Describing APIs for the programmatic access to datasets; *e.g.* by leveraging `dcat:Distribution`, `epos:Webservice` and `hydra:Operation`.
- Strengthening engagement with scientific communities supporting the inclusion of domain specific knowledge; *e.g.* via `skos:ConceptScheme` and `skos:Concept`.
- Enabling user-driven approaches and tagging (via annotations); *e.g.* `epos:annotation`.
- Enabling integrity checks and validation (via SHACL validators).

The latest version of the EPOS-DCAT-AP data model is available online[10] – it includes a UML diagram, ontology definition, shapes graphs (in SHACL), examples and more details. Detailed information can be found also in Appendices A and B. Section 5.3.3.1 provides an overview of EPOS-DCAT-AP and its applications. Following the DCAT philosophy we reused well-known bundles such as Schema.org and the Web Annotation Vocabulary. When reuse was not possible we created extensions in the EPOS namespace.

[10] `https://github.com/epos-eu/EPOS-DCAT-AP/tree/EPOS-DCAT-AP-shapes`

The `WebService` entity has been modelled leveraging Schema.org and the Hydra Core Vocabulary [Lanthaler, 2018] for evolvable Web APIs, they were both introduced in Chapter 3. This allows us to have flexible and fine-grained representations covering the broad EPOS spectrum that includes both global, well-established and community specific standards for web services *e.g.* OGC and FDSN. RDF allows us to include existing domain specific namespaces thus supporting community and user-defined bundles. The `Annotation` entity can be harnessed to enable the collaborative, 'folksonomical' approach – Core Concepts can be enriched with user-driven descriptions and new concepts can be created aggregating, grouping and connecting existing concepts. An important feature is the support for integrity and validation embedded in the representation. This is achieved via the Shapes Constraint Language (SHACL).

It is worth mentioning that the availability of tools that allow representational translation, such as X3ML by FORTH [Minadakis et al., 2015], might make the choice of a specific representation less sensitive. For instance, a mapping of EPOS-DCAT-AP to CERIF is available [Theodoridou et al., 2019]. Where needed, multiple representations might coexist without affecting the conceptual definitions of the CC.

5.3.3.1 EPOS-DCAT-AP and examples of its application

In this section we present some details of EPOS-DCAT-AP including a high-level UML class diagram of the EPOS-DCAT-AP model and examples of encodings in the RDF/Turtle notation. Figure 5.6 shows details of the extensions built on top of DCAT-AP v1.1 [European Commission, 2015a]. A complete UML diagram of EPOS-DCAT-AP is presented in the Appendix A.

The additional classes introduced are represented in yellow. They allow us to address the specific requirements of the EPOS community. In particular, they enable the description of additional concepts beyond datasets (the main focus of DCAT). For instance, `Service` and `WebService` allow the mapping of important community assets. Such concepts can be included in a catalogue with: $Catalog \xrightarrow{\text{epos:resource}} Resource$ as illustrated in Listing 5.1. As such, they are specialisations of a `Resource` that is a generic concept extending the range of catalogued types: *Service is_a Resource* and *WebService is_a Resource*. A similar feature has been recently introduced in a revised version of DCAT that includes `dcat:Resource` [Beltran et al., 2018]. Listing 5.1 provides an example of a catalogue including different resources.

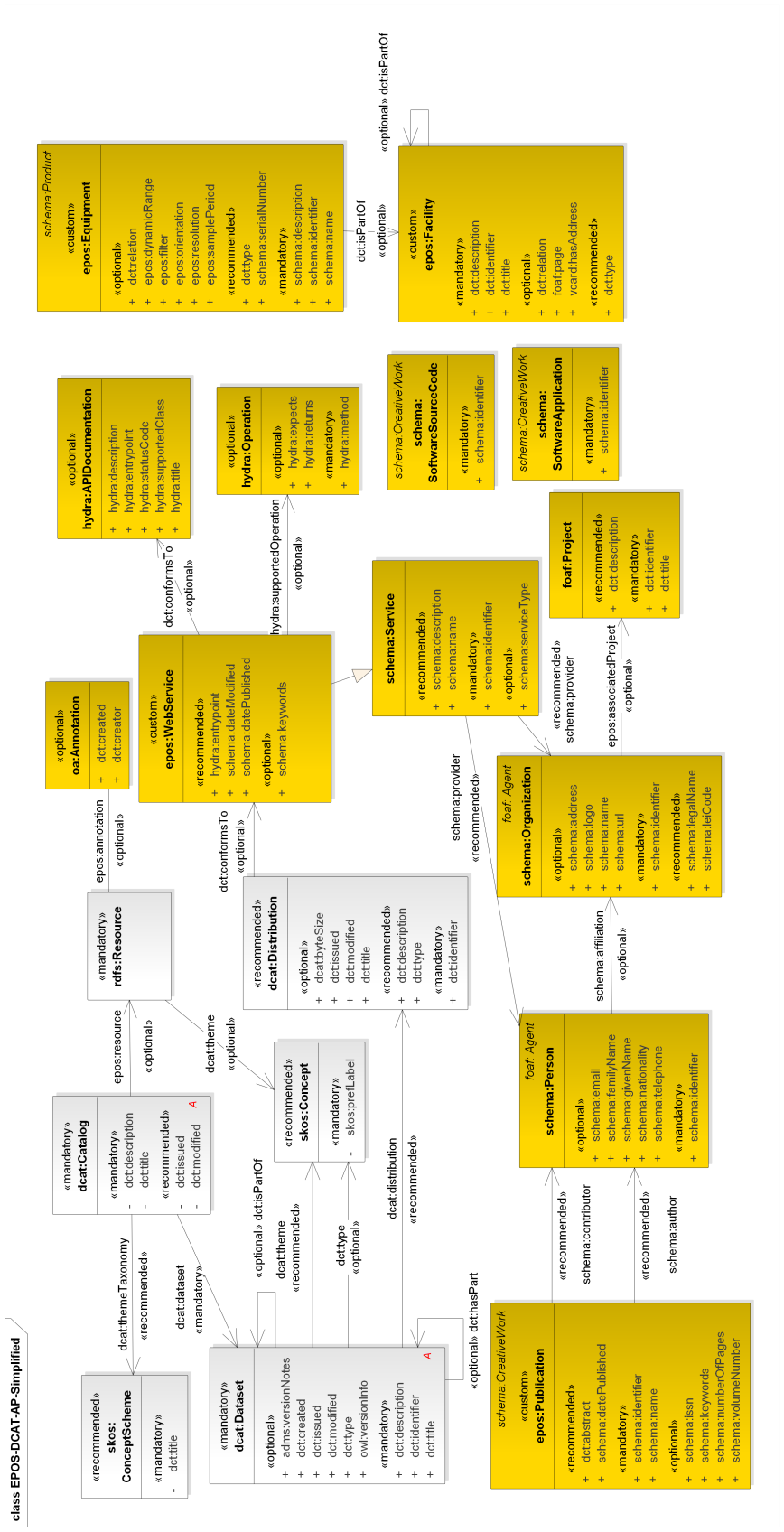


Figure 5.6: Simplified UML model of EPOS-DCAT-AP. We extended DCAT-AP v1.1 with additional classes (in yellow) and their relationships in order to fulfill the EPOS requirements. Stereotypes indicate 'mandatory', 'recommended' and 'optional' elements. Similarly, 'custom' indicates additional classes defined in the epos namespace.

Listing 5.1: It shows the start of the definition of the conceptual space, `ConceptScheme`, for EPOS, `Epos`. The established namespaces from which terms are imported are defined. The concept `catalogID` is then introduced and the first five attributes of its elements to hold metadata are defined. Others are omitted. Their resources, `resource`, are then defined, but for clarity their details are omitted. A resource in this context is an asset relevant for EPOS, *e.g.* a Facility, an Equipment and a Web Service. Format is RDF/Turtle.

```
@prefix epos: <http://www.epos-eu.org/epos-dcat-ap#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
## A scheme that includes EPOS concepts
   <epos:Epos> a skos:ConceptScheme;
        dct:title "EPOS concepts"@en;
  dct:description "It contains the concepts of the EPOS domain"@en; .
## A catalogue that collects EPOS assets
   <catalogID>  a dcat:Catalog;
  dct:title "EPOS Metadata Catalogue"@en;
  dct:description "A catalogue that represents the EPOS CC"@en;
  ## A skos taxonomy of the domain specific concepts in EPOS
  dcat:themeTaxonomy <epos:Epos>;
  ## Including datasets and data collections
  dcat:dataset <datasetID>;
  dcat:dataset <datasetCollectionID>;
  ...
  ## Including additional assets
  epos:resource <equipmentID>;
  epos:resource <facilityID>;
  epos:resource <webserviceID>
  ... .
```

Webservices are very important in EPOS as they provide programmatic access to a variety of datasets and resources. In EPOS-DCAT-AP we are able to describe such a programmatic access by harnessing the `dcat:Distribution` class – "*Represents a specific available form of a dataset*" [Maali and Erickson, 2014] – with a specific application of its relationships: *Distribution* $\xrightarrow{\text{dct:conformsTo}}$ *WebService* and *Distribution* $\xrightarrow{\text{dcat:accessURL}}$ *Operation*. Such an application is defined as a SHACL graph in Listing 5.2 – the complete SHACL description of EPOS-DCAT-AP is available in the Appendix B.

Listing 5.2: A shapes graph that specifies the application of a Distribution in EPOS-DCAT-AP. *E.g.* it extends the scope in order to access a data service. Format is RDF/Turtle.

```
@prefix epos: <http://www.epos-eu.org/epos-dcat-ap#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix hydra: <http://www.w3.org/ns/hydra/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dct: <http://purl.org/dc/terms/> .
epos:DistributionShape a sh:NodeShape ;
sh:targetClass dcat:Distribution;
####
# Distribution mandatory properties
###
sh:property [
      sh:path dcat:accessURL ;
## it can be a URL to a generic resource
     sh:or (
 [ sh:class rdfs:Resource ; ]
 ## or to a specific web service method/operation
 [sh:class hydra:Operation ;
   ] ) ;
 sh:minCount 1 ;
 ] ;
    ...
####
# Distribution Recommended properties
###
sh:property [
sh:path dct:conformsTo ;
     sh:or (
## it can conform to an existing standard
        [sh:class dct:Standard ; ]
## or to a webservice specification
        [sh:class epos:WebService ;
        ] ) ;
       ] ;
...
```

Instruments can be described with the concept `Equipment` and linked to a specific `Facility` *Equipment* $\xrightarrow{\text{dct:isPartOf}}$ *Facility*. Listing 5.3 provides an example of a seismic network described as an `epos:Facility` with a seismic station (`epos:Equipment`) and a seismic stream (channel) (`epos:Equipment`), such a representation is obtained by mapping a seismological standard: StationXML[11].

Listing 5.3: Example of mappings of Facility and Equipment. A classification of seismological concepts is defined. It includes the concepts `SeismicNetwork` and `SeismicStation`. Then instances of concepts, `Facility` and `Equipment`, are defined with some of their attributes. Others are omitted for clarity. Format is RDF/Turtle.

```
## A classification of seismological concepts
    <epos:Seismology> a skos:ConceptScheme;
        dct:title "Seismology"@en;
        dct:description "It contains the concepts of the Seismology domain"@en; .
## Defining the concept Seismic Network
    <epos:SeismicNetwork> a skos:Concept ;
        skos:definition "Collection of seismic stations in a seismic network";
        skos:inScheme <Seismology> ;
        skos:prefLabel "Seismic Network" .
## Defining the concept Seismic Station
    <epos:SeismicStation> a skos:Concept ;
        skos:definition "A station for recording oscillations of the Earth's surface";
        skos:inScheme <Seismology> ;
        skos:prefLabel "Seismic Station"
        skos:altLabel "Seismometer".
## Describing a seismic network (NL) as a Facility
    <EPOS/ORFEUS/EIDA/ODC/NL> a epos:Facility ;
        dct:description "Netherlands Seismic and Acoustic Network";
        ##Seismological networks follow the FDSN recommendation to adopt
        ##DOIs (www.fdsn.org/services/doi/)
        dct:identifier <doi.org/10.21944/e970fd34-23b9-3411-b366-e4f72877d2c5> ;
        dct:title "Seismic Network NL";
        dcat:contactPoint <ContactID> ;
        ##The concept associated with this Equipment
        dcat:theme <SeismicNetwork>;
        ... .
## Describing a seismic station as an Equipment
    <EPOS/ORFEUS/EIDA/ODC/NL.HGN> a epos:Equipment ;
        dct:description "Broadband Seismic Station HEIMANSGROEVE, NETHERLANDS " ;
        dct:identifier <EPOS/ORFEUS/EIDA/ODC/NL.HGN> ;
        ## Location
        dct:spatial [ a dct:Location ;
            locn:geometry "POINT(50.764 5.9317 135.0)" ] ;
        dct:title "Seismic Station NL.HGN";
```

[11] www.fdsn.org/xml/station/

```
        ##The concept associated with this Equipment
        dcat:theme <SeismicStation>;
        ## This station belongs to the NL network
        dct:isPartOf <EPOS/ORFEUS/EIDA/ODC/NL>;
        ... .
## A seismic stream belonging to a station
    </EPOS/ORFEUS/EIDA/ODC/NL.HGN.02.BHZ> a epos:Equipment ;
        dct:description "Seismic stream recording ground motion";
        dct:identifier <EPOS/ORFEUS/EIDA/ODC/NL.HGN.02.BHZ>;
        ## This stream belongs to the NL.HGN station
        dct:isPartOf <EPOS/ORFEUS/EIDA/ODC/NL.HGN> ;
        dct:spatial [ a dct:Location ;
            locn:geometry "POINT(50.764 5.9317 135.0)"] ;
        dct:title "Seismic Stream NL.HGN.02.BHZ";
        epos:orientation "0.0/-90.0";
        epos:samplePeriod "0.025";
        ... .
```

The Listing 5.4 shows an example of classification using SKOS. It can be used to describe domain specific concepts (*e.g.* Seismic Waveform) which can be associated with the EPOS Core Concepts *e.g.* Dataset, Webservice: $Resource \xrightarrow{\text{dcat:theme}} Concept$.

Listing 5.4: Example of classification using SKOS. It groups knowledge in concept schemes, `ConceptScheme`. Here we see a few members, `Concept`, be gathered under the `Seismology` theme's heading, and then a group of concepts being gathered under the `VolcanoObservations` heading, with one concept, `SeismicWaveform` shared. Format is RDF/Turtle.

```
@prefix epos: <http://www.epos-eu.org/epos-dcat-ap#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
##### Example of a classification of domain specific concepts to be associated with the EPOS CC
#####
## Communities can define and manage their sets of concepts in a concept scheme
<epos:Seismology> a skos:ConceptScheme;
        dct:title "Seismology"@en;
  dct:description "It contains the concepts of the Seismology domain"@en; .
## Defining the concept SeismicWaveform with multi-lingual support
<epos:SeismicWaveform> a skos:Concept;
  skos:definition "Measurement of the dynamic displacement of the Earth"@en;
  skos:inScheme <epos:Seismology>;
  skos:prefLabel "Seismic waveform"@en;
  skos:prefLabel "Forma d'onda sismica"@it;
  #can be used by applications for text-based indexing/search (e.g via a web interface)
  skos:hiddenLabel "seismic_waveform"@en;
        skos:hiddenLabel "MSEED"@en; .
## Another seismological concept
```

```
<epos:SeismicHazardMap> a skos:Concept;
  skos:definition "A map that shows the hazard associated with potential earthquakes in a particular
     area";
  skos:inScheme <epos:Seismology>;
  skos:prefLabel "Seismic hazard map"@en ;
  skos:altLabel "Seismological hazard map"@en; .
<epos:VolcanoObservations> a skos:ConceptScheme;
        dct:title "VolcanoObservations"@en;
  dct:description "It contains the concepts of the Volcano Observations"@en;  .
<epos:GeochemicalData> a skos:Concept;
  skos:definition "It refers to the types of geochemical  ...."@en;
  skos:inScheme <epos:VolcanoObservations>;
  skos:prefLabel "Geochemical Data"@en;
        skos:altLabel "Geochemistry"@en; .
## SeismicWaveform belongs to more that one concept schemes, i.e. it is a shared concept
<epos:SeismicWaveform> a skos:Concept;
  skos:definition "Measurement of the dynamic displacement of the Earth"@en;
  skos:inScheme <epos:VolcanoObservations>;
  skos:prefLabel "Seismic waveform"@en;
  skos:altLabel "Seismology"@en; .
## Importing an existing ontology. Communities who already invested in the definition
## of formalised knowledge can retain their investments.
<CommunityOntology> a owl:Ontology, skos:ConceptScheme .
```

EPOS-DCAT-AP has been conceived for the solid-Earth sciences community, nevertheless it fulfils requirements in common with many Research Infrastructures and it can be applied in broader contexts [Trani et al., 2018c]. For more details and examples we refer readers to the Appendices A and B and to the online documentation[12].

5.3.4 Population of the EPOS Canonical Core

Once the EPOS Core Concepts have been identified and agreed, and an appropriate representation chosen, the next step is the population of the CC with real entities from the designated communities. This demands close interaction and collaboration between domain and metadata experts. Ultimately, population needs to be a process that is automated as far as possible. But this requires preparatory work. First experts need to agree the data sources for each concept. They then need to develop import-transformations and protocols. These may stimulate changes at sources and in the CC. Once validated, the parties involved need to agree to sustain the relationships and then an automated process can be coded and run whenever necessary.

[12] https://github.com/epos-eu/EPOS-DCAT-AP/

To kick off the population process dedicated meetings and workshops were organised targeting the EPOS communities. Documentation, training material, demos and webinars were delivered prior to the face-to-face events in order to inform and prepare the communities for the effort required. This needed to develop the motivation and stimulate the commitment of effort. Moreover, collaborative tools such as wiki and shared repositories have been set up to collect the inputs and feedback from the communities and share documentation and results. To achieve the preliminary population of each community's bundle into the EPOS CC, the communities had to map their resources to the corresponding concepts of the EPOS CC with support from the EPOS-DCAT-AP experts. Due to the scale and complexity of this process the mapping has been carried out in stages prioritising specific entities and adopting in an initial phase a simplified XML representation. Table 5.3 shows the population of the initial entities.

Entity name	**Number of instances**
Person	86
Organisation	32
WebService	74

Table 5.3: Number of instances of the prioritised entities after the initial (manual) population and validation. In the next stages of the population process (automated) a substantial increase of the number of instances is expected. *E.g.* Person is expected to grow at least by a factor of 100, Organisation by a factor of 10 (nearly 260 have been surveyed). The number of instances of Web Service will likely stay in the same order of magnitude and grow at a slower pace. However, in this case it is important to note that the populations made available indirectly by those services are very large.

During the initial population each community uploaded EPOS-DCAT-AP XML compliant files on the EPOS GitHub repository[12]. Those files were successively manually curated and validated. This manual process, albeit costly, helped by testing the knowledge collection process and by validating the model chosen for the representation. Moreover, it showed an active engagement and participation of the communities who provided useful contributions and feedback. In this phase it has been particularly challenging to keep the alignment of the population with the ongoing refinements of the representation of the core (*i.e.* EPOS-DCAT-AP). This dynamic situation some-

times introduced issues for the communities, in Section 5.4 we evaluate some of those issues with the challenges addressed. According to the principles described in Section 5.1.4 one of the goals of the population process is to specify the type of ingestion strategy (1) for each entity (*i.e.* harvesting vs brokering). Therefore the communities have to indicate the requirements and accrual policies associated with their entities – EPOS-DCAT-AP supports this information.

In an operational system the processes of mapping, harvesting and/or brokering of entities are typically delegated to automated methods and tools as illustrated in Chapter 3 (3.4.5). To perform the population of the community bundles on a larger scale we devised an architecture with automated components shown in Figure 5.7. Transformations, convertors and parsers are used to extract the information required by the CC directly from community bundles. SHACL validators (3.2.3) help discriminate the admission of entities into the CC and debug representation errors. Nevertheless, those technical solutions depend on agreements between the sources of information and the CC, commitments to fulfil those agreements and good behaviour – these are essential to ensure that consistent information is delivered.
Examples of mappings of community bundles are available in the EPOS-DCAT-AP GitHub[13]. In Chapter 6 we return on these aspects in Section 6.2.4.

5.4 Initial Evaluation

In this section we provide a preliminary evaluation of the application of our methodology in EPOS. This can be seen as a baseline to refer to when we address the evaluation in Chapter 6. However, it is important to note that a number of reasons make a complete evaluation of the impact infeasible at this stage. The nature and scope of the issues addressed in this research require a longer time scale to be effectively measured. There are individual and organisational aspects that influence adoption and uptake. Those are critical within a single organisation and become much harder in multi-organisational and multi-disciplinary contexts. We target sharing behaviours and working practices that require time to assimilate novel elements. EPOS is currently in its implementation phase [Cocco, 2018], for a more complete assessment evaluations ought to be repeated when it is transitioning to its operational phase. These should then be repeated peri-

[13] https://github.com/epos-eu/EPOS-DCAT-AP/tree/EPOS-DCAT-AP-shapes/examples

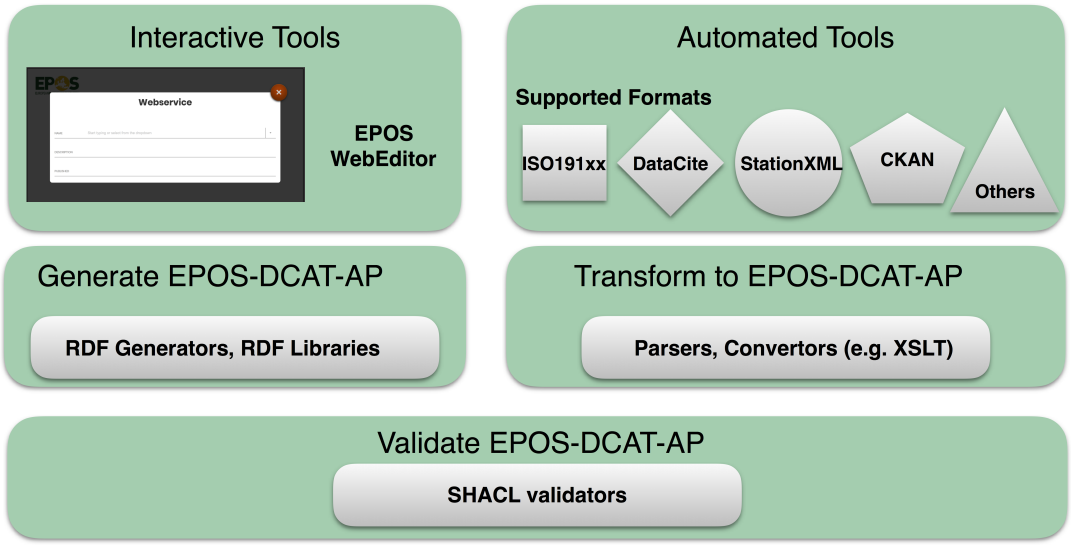


Figure 5.7: Supporting the population process with automated tools. The diagram indicates the components involved in the population. The ingestion can be performed in an interactive way with the help of a Web Metadata Editor [Riccardo Rabissoni, 2018]. Alternatively a programmatic way is available for a number of supported formats. The converters for each domain specific format are built either reusing existing tools in the case of standards such as ISO191xx, DataCite; or in collaboration with the communities (*e.g.* StationXML). The validation is performed using SHACL validators and the shapes defined for the EPOS model.

odically to detect trends. We build on similar approaches that exploit catalogues and agreed canonical forms in the seismological domain [Trani et al., 2017]. That experience, described in Chapter 4, provided us with useful evidence of benefits and take up although it has been applied in a simpler context. In this section we report an initial assessment of our work by highlighting some of the challenges encountered and addressed while engaging with the EPOS communities. A broader evaluation of our approach that includes further iterations is provided in the next chapter.

In a recent meeting[14] (March 2018) we asked various key representatives of the EPOS communities, including developers, technical contacts of diverse Thematic Core Services (TCS), domain and metadata experts, leaders and coordinators to provide their feedback about the EPOS-DCAT-AP model. Table 5.4 contains the questions

[14] www.epos-ip.org/events/epos-implementation-and-validation-workshop-lisbon%2Dportugal12-14-march-2018

and the responses from the participants. Out of approximately 25 participants, who were involved in developing shared knowledge for EPOS, 13 returned responses. In general their responses show a very promising consensus. There are some aspects to improve but it is clear that since the initial presentation of the model to a meeting of four (out of ten) EPOS themes[15] in May 2017 (10 months earlier), a substantial awareness and understanding had been achieved. The responses to Question 1 show an almost unanimous consensus about the usefulness of the model as a means to facilitate the collection and exchange of domain knowledge.

Table 5.4: First evaluation survey about EPOS-DCAT-AP

Question	# of responses
1: Would the introduction of EPOS-DCAT-AP facilitate the collection and exchange of domain knowledge for the EPOS-ICS?	13
2: Please identify limitations in the proposed EPOS-DCAT-AP. As many as you wish. Where you have suggestions as to how they should be addressed please feel free to make them.	13
3: Are there other contexts that you or your organisation work in where the approach leading to EPOS-DCAT-AP would be useful? Please identify them.	10

Question 2 provides interesting feedback about the perceived limitations of the model. Participants report a number of issues which in some cases have been collected in the EPOS-DCAT-AP GitHub[16]. However, most of those issues are related to the initial XML version of the model. In the current RDF version they have been solved. For instance, a better description of `WebService`, building on Schema.org and the Hydra Vocabulary, has been introduced to address previous limitations. Concerns about the population strategy and complexity have been addressed by introducing automated tools. A broader set of roles to better support attribution information has been suggested as a possible improvement. We acknowledge the importance of such a requirement and considered to include elements from the PROV vocabulary [Lebo et al., 2013] in order to provide a broader structured provenance information. However, we decided to postpone such a feature to later versions and proposed an intermediate

[15] `https://www.epos-ip.org/milano-and-rome-epos-harmonization-meetings`

[16] `https://github.com/epos-eu/EPOS-DCAT-AP/issues`

solutions within the current model. One of the answers points out the importance of an agreed strategy for the identification of resources, a feature strongly promoted by EPOS-DCAT-AP – "*The biggest problem (outside of the model) is assigning identifiers to entities and make sure these are consistent [...] We need to agree on rules to set these identifiers and collect them in a (single) repository*". This is an example of an engineering aspect that depends on shared agreements. It shows an increased awareness achieved in the communities about important issues and how our approach stimulated the thinking towards a common shared solution.

Finally, the responses to Question 3 offer the following reflections. A couple of answers identify interesting application contexts. One is positive but shows caution – "*[...] If the EPOS extensions are accepted back into future DCAT standards, this may make implementing EPOS DCAT more attractive*". As we show in Chapter 6 there is evidence of our collaboration influencing the DCAT development. The remaining answers are more reluctant. For sure at this stage there is still not enough knowledge and trust that would warrant migration from established practices. This is reasonable and in line with the expectations. As already mentioned the introduction of novel elements requires time. Also, local contexts quite often develop solutions tailored to specific needs, the complexity associated with generalisations required in broader contexts can be perceived as an unnecessary overhead. In any case it would be useful to repeat this evaluation when more experience has been acquired and to assess the benefits delivered.

To conclude this analysis, we highlight some key outcomes: the collaborative interaction has been very successful and productive, it allowed us to collect feedback and improve many aspects (*e.g.* of the conceptual model, the procedures for obtaining agreement, the representation and the population processes) in order to better support communities' requirements. It encouraged us to think about issues previously unanticipated and developed a common vocabulary and improved the understanding of concepts. This suggests we have a foundation and *modus operandi* for sustainable incremental progress.

5.5 Conclusions and discussion

The concept of Information-Powered Collaborations (IPC) introduced in this thesis is an abstraction that captures the complex dynamics of a modern research context that depends on multi-organisational, multi-disciplinary, multi-national collaboration with increasing complexity and scale. In this chapter we proposed the formation of an explicit Canonical Core (CC) as the foundation for information sharing and a framework that partitions the complex task of agreeing and maintaining a consistent set of shared Core Concepts to sustain interdisciplinary collaboration. That set has three independent aspects: conceptual definition, representation and population. We have demonstrated how such a framework facilitates the construction and evolution of the information space underpinning an IPC by enabling successive refinements of the three aspects. For instance, communities who are mainly interested in having their entities (*e.g.* data, services and methods) available in the CC will focus on the population. Those developing automated methods might find the current representation is missing aspects needed and therefore require additions to the representation of the CC. Similarly, someone interested in extending high-level goals might enrich the set of Core Concepts. Thanks to our framework those issues can be addressed independently and progressively, thereby exploiting a separation of concerns. Another important advantage of our framework is that it supports innovation, experiments and heterogeneity. It enables the retention of valued working practices in the Boundary Regions until it is beneficial to transition them into the core, thereby minimising disruption, avoiding constraints and pursuing continuous incremental adoption. Furthermore, it fosters more efficient communication and progressively negotiated agreements between the stakeholders by partitioning the dialogue. As communication is particularly challenging in multi-disciplinary, multi-cultural environments the presented framework provides a significant advance that has been tested in EPOS. We will continue with this approach in EPOS. In particular we plan to:

1. maintain the current set of Core Concepts evolving the Canonical Core when required by new requirements and use cases;

2. further develop and refine the EPOS-DCAT-AP representation, by strengthening our collaboration with the W3C by working with the DXWG [W3C-DXWG, 2018] in order to make it available for other communities;

3. provide tools leveraging existing components to better support the designated communities in the automated population of their entities. For instance, by means of: graphical interfaces, convertors, mapping services, workflows *etc.*; and

4. work on the integration of annotation management tools such as EUDAT B2Note[17] to further exploit the collaborative approach.

In the next chapter we illustrate how the most of these points have been addressed.

Establishing collaborative knowledge to achieve holistic integration and semantic interoperability is an extremely complex task of wide interest that requires alignment of technical, organisational and cultural factors. In order to succeed in this endeavour implications and issues ought to be recognised and addressed effectively, stakeholders acknowledged and good behaviour properly rewarded, *e.g.* by promoting evidence of enhanced scientific results and increasing return on investments. Accommodating local diversity while encouraging migration towards and engagement with the core is essential for sustaining effective collaboration. Although a long way still remains along this path, we believe that the set of principles, the philosophy and the approach proposed are important initial steps.

[17] `https://b2note.bsc.es/`

Chapter 6

Evaluating the methodology for empowering IPC

The goal of this chapter is to provide readers with an evaluation of the methodology presented in this thesis. The criteria underpinning such an evaluation are the following:

1. Feasibility – our approach's degree of being easily or conveniently applied.
2. Utility – the quality of our approach to fulfil the addressed requirements.
3. Usability – the extent to which our approach can be adopted by the target communities.
4. Sustainability – the quality of our approach to be effectively exploited over time by the target communities.

In the limited time-frame of this research we were not able to perform a complete evaluation of all those criteria. In Chapter 5 (5.4) we discussed challenges and issues, which are inherent to the targeted scale and objectives, that prevented us from achieving a comprehensive assessment of benefits and impact. We address socio-technical aspects and behaviours that require repeated and continued assessments in order to identify trends and collect evidence of long-term impact.

These reasons suggested us to design an evaluation approach and devise a framework that can be harnessed consistently and regularly in future measurements. We recognise that in the frame of this research we are able to lay only the foundations for

such a framework. For a full characterisation and formalisation further investigations will be needed. It will benefit from and build on the results presented here.

We start by analysing the tractable context described in Chapter 4, *i.e.* WFCatalog. This provides us with interesting insights thanks to its relative maturity.

6.1 A retrospective on WFCatalog

We draw an assessment of WFCatalog describing the current (October 2018) status and its achievements since the operational launch. Conclusions, based on the assessment criteria presented in the introduction of this chapter, are drawn in Section 6.1.4.

The first operational deployment of WFCatalog was performed at the Orfeus Data Centre in Nov 2016. It then progressed gradually to be adopted in all the ten EIDA data centres and achieved completed European adoption in the first half of 2018. Since 2018 WFCatalog is part of the official ORFEUS EIDA service portfolio[1].

The information acquired in the significant time elapsed since the first deployment allows us to perform a quantitative analysis of the progress achieved so far. We are particularly interested in understanding the impact of such a service and its focused core on the user communities and the influences in their communication and information exchange processes. We carry out the evaluation by applying the same approach that we have adopted throughout this thesis, *i.e.* by addressing the three dimensions separately: conceptual definition, representation and population.

6.1.1 Conceptual definition

We remind the readers about the main objective of the conceptual definition: it targets the information space where common concepts are defined, discussed and shared. Such a conceptual space in WFCatalog encompasses the definition of seismic waveform characteristics and features (*e.g.* quality metrics). For instance, it deals with agreements on common seismic waveform characteristics and the socio-political framework supporting such decision-making processes. In Chapter 4 we discussed the challenges associated with such a space. Also, we presented the plans for future exploitation and extensions that were set at that time (2016). One of our primary goals

[1] https://www.orfeus-eu.org/data/eida/webservices/

was to promote uptake beyond Europe and to establish WFCatalog and its concepts globally by targeting FDSN. We have seen how the agreement process in Europe yielded important results and a wide consensus among EIDA data centres was reached. We also discussed how such a process had a different pace at FDSN and motivated this with the broader scale and the loose coupling of the stakeholders (in terms of priorities and levels of commitment). After nearly two years those discussions are still ongoing. Although some definitions have been shared and applied in similar applications [Casey et al., 2018] and the importance of the concepts and the need for such a service are well received and recognised [Vecsey, 2018; Schuh et al., 2018], a misalignment of priorities slows down the agreement process. This confirms that the socio-political aspects are predominant. Also, it reinforces the motivation to preserve such community investments when integrating them into wider cross-disciplinary, holistic systems.

Another aspect concerns the evolution of the set of concepts, *i.e.* features and quality metrics. Some of them were already identified, namely Power Spectral Density (PSD) functions. As their definitions are quite broadly recognised and accepted in the seismological community, it was possible to implement them in a relatively short time. We anticipated this in Chapter 4. However, some technical details remained to be discussed before the eventual roll-out in a production version. Examples are provided in the next section as they target the representation dimension.

There is evidence of impact at conceptual level as WFCatalog helped address FAIR-ness of seismic waveform and exchange them in broader contexts beyond seismology [Koymans et al., 2018; Atkinson, 2018; Trani and the EPOS-ORFEUS-CC Team, 2018; Magagna et al., 2018]. Also, WFCatalog has been registered as one of the resources contributed to the EPOS catalogue[2] – it is one of the recognised community bundles.

We conclude from observing these repeated uses that the adoption of WFCatalog's set of core concepts stimulated knowledge exchange and promoted engagement of the seismological community and beyond. Nevertheless, even for such focused core a longer observation time will be required in order to have a better picture about utility and sustainability – the latter is of particular interest for us.

In the next section we analyse progress achieved along the second dimension: representation.

[2] https://www.epos-ip.org/tcs/seismology/data-services/list-services

6.1.2 Representation

The representation dimension addresses the way concepts are organised and structured and their encodings. In the context of WFCatalog representation entails, for instance, its data model and exchange formats. In this respect, one of the goals initially identified was to improve interoperability with broader communities and enable interoperation by introducing Dublin Core metadata elements. This was scheduled and discussed in a dedicated team. An extension of WFCatalog was then implemented by one of the EIDA partners, INGV, and is now (October 2018) available[3]. It includes the following Dublin Core metadata elements: Identifier, Title, Subject, Creator, Contributor, Publisher, Type, Format, Date, Coverage, Available, DateAccepted and isPartOf. This was a nice example of fruitful interaction and collaboration fostered by WFCatalog – partitioning the discourse into conceptual and representational aspects introduced a path to tackle and organise collaborative work.

Such an extension has been adopted to support an application in the context of the European Open Science Cloud hub project (EOSC-hub[4]). In particular, WFCatalog will be harnessed to enable discovery of seismological waveforms and their staging onto computational infrastructures [Trani and the EPOS-ORFEUS-CC Team, 2018]. That application demonstrates how, thanks to adequate descriptions, seismological waveform data can be shared and exploited across domains. At present (October 2018) Dublin Core is still an extension that would require further agreements before it could be included in a next production version of WFCatalog. This would agree about additional parameters required in the WebAPI and changes to the serialisation format.

Another application of the WFCatalog metadata was piloted in the EUDAT project[5]. WFCatalog's data model was exploited to enable users to compute seismic waveforms and generate standard descriptions and quality metrics local to their data holdings. The adoption of the WFCatalog descriptions enables the exchange of consistent results, thus enforcing the importance of shared definitions and a standard representation.

Additional quality metrics, *i.e.* Probability Spectral Densities (PSDs), are available as an extension which is running at the ORFEUS Data Centre (ODC). To be included

[3] https://github.com/massimo1962/WF-DO_dc_airods

[4] https://eosc-hub.eu/

[5] https://github.com/EUDAT-GEF/GEF/wiki/Seismological-Use-Case

in an official version such an extension requires adaptations of the `WFCatalog` data model as well as of the WebAPI. A description of the WebAPI based on the OpenAPI/Swagger[6] standard is available[7] since Dec 2017 – it provides a better support for the automated generation of clients.

Also, we considered a serialisation of the `WFCatalog` metadata as Linked Data, in particular as JSON-LD. This feature has been implemented in a prototype and allocated to a future revision.

We conclude the assessment of the representation dimension by recognising significant progress achieved in the application of the data model underpinning `WFCatalog`. It should be noticed that representation challenges could be addressed and piloted quite rapidly. Nevertheless, they depend on broad agreements to be finally adopted and rolled-out into production. Those progress at a slower pace.

The consideration of extensions to `WFCatalog` seems to indicate: a) that there is an active community using it; and b) that as that community explores and adopts extensions, it shows utility and sustainability.

6.1.3 Population

The population dimension targets the management of instances of concepts and their relationships. In this case, for instance, it provides indications of the variation of scale (*e.g.* number of deployments, users) and volumes involved (*e.g.* number and types of metrics). In this section we provide an update about those aspects.

Since its launch `WFCatalog` has been deployed in ten data centres that operate it as a primary service. To provide an estimate of the growth of the populations that are made available, we focus on the ORFEUS Data Centre production instance – a summary of the population statistics is available in Table 6.1. We notice a substantial growth of the volumes of metadata instances which is directly related to the size of the underpinning seismic waveform data.

At the same time the still quite low usage of the service stands out. This might be due to an inherent community's inertia to change. For instance, the establishment and uptake of the FDSN services took several years. Also, a misalignment of the status

[6] `https://swagger.io/`

[7] `https://www.orfeus-eu.org/swagger/dist/index.html?url=https://www.orfeus-eu.org/data/eida/webservices/wfcatalog/wfcatalog.yaml`

Item	Launch (Nov 2016)		Current (October 2018)	
Raw seismic waveform	Files	∼4M	Files	∼7M
	Size TB	∼15	Size TB	∼50
Daily stream collection	Count	∼4M	Count	∼7M
	Size GB	1.22	Size GB	7.7
Continuous segments collection	Count	∼400M	Count	∼427M
	Size GB	85	Size GB	∼93
User requests	Count	NA	Count	∼30K

Table 6.1: Population and usage statistics of WFCatalog operated at the ORFEUS Data Centre

of deployments in the data centres of the EIDA federation slowed down outreach and promotional campaigns. For a more complete picture and a realistic assessment, future evaluations will be required that should take into account other EIDA nodes as well.

In conclusion, we showed how the evolution of the population dimension progresses at a faster pace. The scale and volumes involved are much larger when considering the whole of EIDA. There, an established organisational framework ensures that maintenance and operation are fulfilled by shared commitments and agreements for each service in the EIDA portfolio. This guarantees the sustained contribution of new WFCatalog populations by EIDA data centres.

6.1.4 Summary

We presented a summary of the progress of WFCatalog in its short operational timespan. There is evidence of impact and influence within the seismological community and beyond. It demonstrates the utility and feasibility of such a service. However, here the focus is its relation with our methodology.

We described how the definition, representation and population of an agreed set of concepts, albeit of limited scope, can foster and drive collaboration. WFCatalog can be seen as an example of a focused community bundle. Thanks to these features it can contribute to the formation of broader cores *e.g.* the EPOS Canonical Core.

In conclusion, the formalisation of community agreements in shared definitions as metadata representations maintained in catalogues can provide several advantages:

- it can facilitate communication and exchange of information by means of understood, canonical forms;
- it can establish a common and consistent 'language' that promotes improved quality of scientific results;
- it can build trust in results that can be consistently exchanged; and
- it can stimulate new thinking and discussion about novel ideas that drive innovation.

As we showed in the previous sections also in a tractable focused context separating concerns into three dimensions (Conceptual definition, Representation, Population) is a valuable tool both for design and analysis. In the next section we exploit such a tool to address the evaluation of the more challenging context where we applied our methodology: EPOS.

6.2 Evaluating the establishment of the EPOS Core Concepts

After the analysis of `WFCatalog` we move to the more challenging EPOS context. In this case we have a different level of maturity and a much broader focus and larger scale. Any evaluation of our approach in such a large, heterogeneous, distributed research infrastructure is very challenging given the time-span available. Nonetheless, we show that we are able to provide initial evidence by embracing multiple viewpoints into a combined assessment strategy with quantitative and qualitative elements.

The quantitative approach enables us to address the first two criteria: feasibility (1) and utility (2). The qualitative approach is exploited to address usability (3). The last criterion, *i.e.* sustainability (4), is difficult to assess at present. However, we present arguments for its estimation by using examples and scenarios. For instance, in Section 6.2.2.1 we propose a process that leverages key aspects of our methodology and discuss its potential benefits to sustain shared vocabularies in EPOS. Future measurements will provide more accurate indications of the validity of this approach.

As in Section 6.1 we tackle the three dimensions (C, R and P) individually. Our quantitative assessments investigate the progress achieved in terms of volumes,

adoption and usage. The qualitative assessments focus on usability which "*does not exist in any absolute sense; it can only be defined with reference to particular contexts*" [Brooke, 1996]. To better characterise such a complex context we considered two different and complementary evaluation methodologies.

The first is based on a surveying technique, thus explicitly targeting perceived usability. Inspired by the the well-known System Usability Scale [US Department of Health and Human Services, 2013; Brooke, 2013] – a method particularly indicated for measuring usability of products and services (*e.g.* web interfaces) – we designed four surveys on different topics:

1. Approach to manage shared knowledge in EPOS
2. EPOS-DCAT-AP
3. Tools for the population of the EPOS Canonical Core
4. Approach to manage shared vocabulary in EPOS

Those surveys were conducted at one[8] of the EPOS meetings (October 2018) where they were supplied to key representatives of scientific communities and technology experts in the order showed above. We present them and analyse their results later in this section. Asking colleagues to complete 4 surveys in succession may have lead to survey fatigue. Also, the order adopted to present the results in this chapter does not reflect the original.

To capture additional aspects from an 'observer' point of view we also considered to employ the 'Critical Incident Technique' (CIT) – "*a method that relies on a set of procedures to collect, content analyze, and classify observations of human behavior*" [Gremler, 2004]. However, that method would require a longer period to collect significant data and to develop effective coding schemes. We started to gather observations and we plan to apply that technique in the future accompanied by new versions of the surveys.

In order to have a better understanding of the analysis illustrated in this section we remind readers about the context that was introduced in Chapter 5. EPOS, the

[8] https://www.epos-ip.org/events/epos%2Dip%2Dit%2Dteam%2Dmeetings%2Dwp6%2Dwp7%2Dbarcelona%2D9%2D11%2Doctober%2D2018

European Research Infrastructure for Solid Earth Sciences, is currently (October 2018) in its last year of its implementation phase – it will transition to operation in 2019. Its long-term governance will be supported by a strong legal framework that was recently granted by the EU with the establishment of the EPOS European Research Infrastructure Consortium [European Commission, 2018].

In Chapter 5 we described how our methodology helped us shaping the construction of a Common Information Space underpinning such a large infrastructure. We illustrated its application by targeting the three dimensions and provided an overview of the initial results. In this section we report status and progress nearly 6 months further into the implementation process.

6.2.1 Introducing our surveying approach

Before digging into the details of our evaluation it is useful to explain the criteria that we followed in the design of our surveys. This will help readers understand context and results. We conclude this section by presenting a first survey entitled: "*Evaluation of the EPOS approach to manage shared knowledge*".

As previously mentioned, our questionnaires were inspired by the SUS. Because of the variety of our context and the elements to be evaluated, which in some cases differ substantially from the typical SUS applications, we performed customisations of the original SUS questionnaire [Brooke, 1996]. As a consequence the interpretation of results might be affected.

A SUS questionnaire is composed of 10 questions. In its original version it contains positive wording questions alternated with negative wording ones (*i.e.* 5 positive and 5 negative). The value associated with each answer ranges from 0 (Strongly disagree) to 5 (Strongly agree). Based on these assumptions its score can be computed as follows:

1. for the positive questions (odd) subtract one from the value of the answer;
2. for the negative questions (even) subtract the value of the answer from five;
3. in this way all values are scaled from 0 (most negative) to 4 (most positive);
4. sum up and multiply to 2.5 in order to convert the scale range from $(0, 40)$ to $(0, 100)$

The benefits (*e.g.* simplicity, effectiveness) and the reliability of SUS are widely acknowledged in the literature [Bangor et al., 2008, 2009; Brooke, 2013]. A slight different version of SUS contains all positive wording questions and it delivers equally valid results [Sauro and Lewis, 2011]. We adopt such an approach in order to minimise errors of interpretations by respondents. This implies that we skip step 2 in the scoring computation procedure and use only step 1 for all the questions. Whilst maintaining a similar approach for the calculation of our score (`Pseudo_SUS`), we performed actions and adjustments that are outlined below:

- each questionnaire was preceded by a presentation about the related topic;
- each questionnaire was introduced by a short explanation text;
- in many cases SUS questions were rephrased (*e.g.* to add more context), in some cases substantially;
- in one case an additional multiple-choice question was introduced; and
- additional free-text comments sections were added after each question (they are not taken into account in the computation of the final score and in this evaluation);

We now introduce our first questionnaire that targets the general approach of our methodology and its application in EPOS. Individual dimensions will be evaluated in the next sections. Figure 6.1 illustrates the questions composing the questionnaire. We can notice the presence of an extra question (Q11) instead of the typical 10. Also, several questions were adapted to fit the purpose of the evaluation that targets a methodology rather than a product. Of about 35 participants, 14 provided answers that are summarised in the figure.

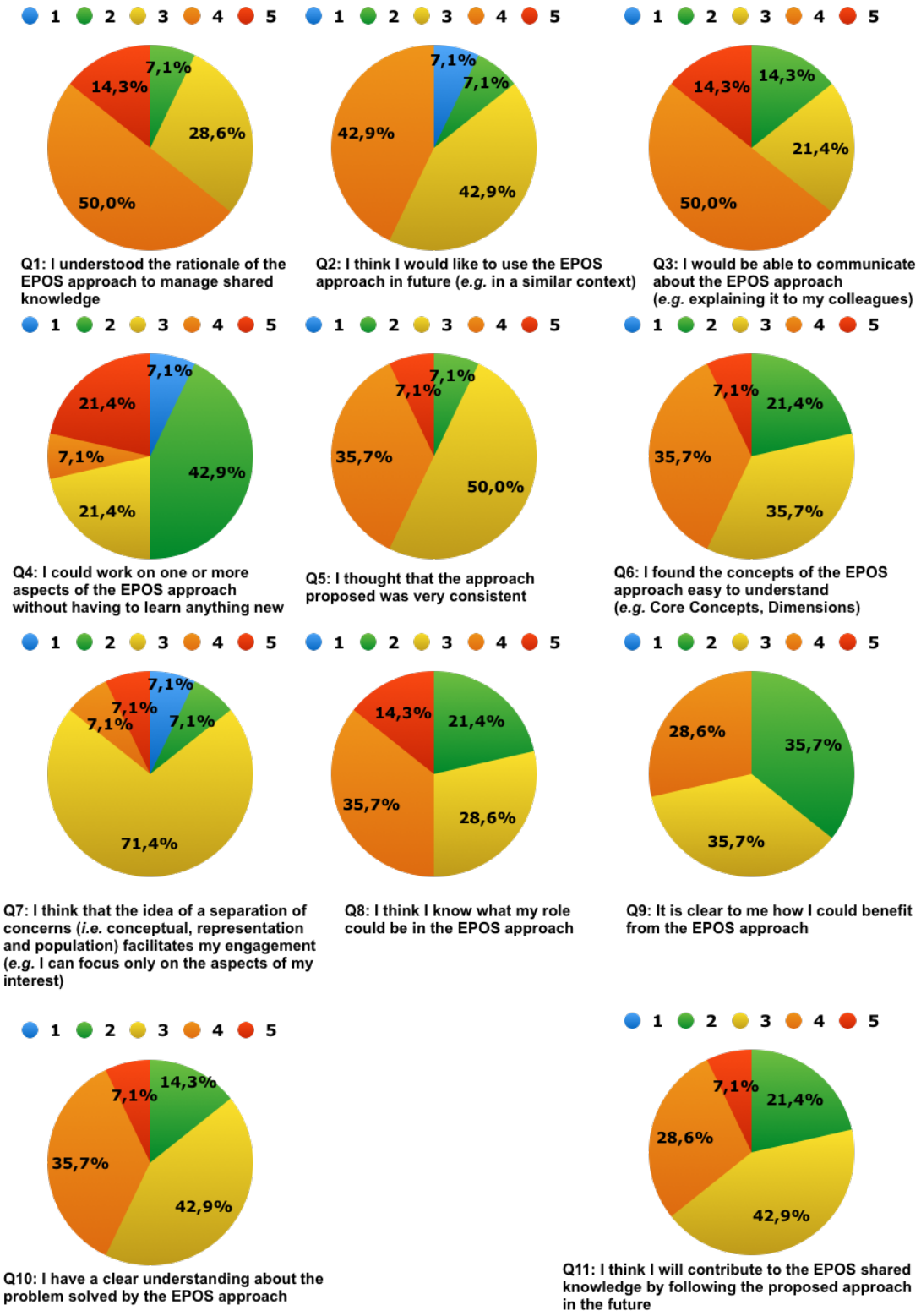


Figure 6.1: Introducing the questionnaire and the results of the survey: Evaluation of the "*EPOS approach to manage shared knowledge*" (herein 'EPOS approach'). The scale of results' values ranges from 'Strongly disagree' (1, blue) to 'Strongly agree' (5, red)

Figure 6.2 shows the average (mean) points scored in each question. The general trend above the average (*i.e.* 3) can be interpreted as a positive indication for the perceived usability of our methodology. In particular, as indicated by some answers (Q1, Q3) the rationale of the approach is well received.

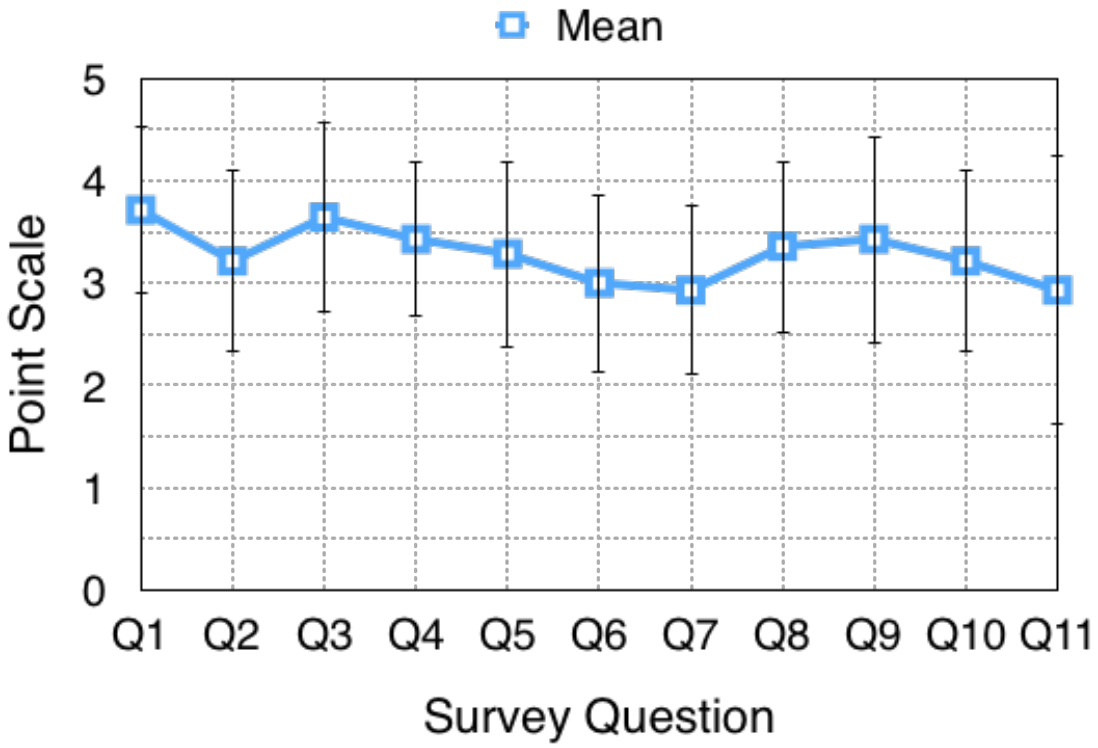


Figure 6.2: Showing the average (mean) points obtained in each question (Q1 – Q11) for 14 responses in every case. Error bars refer to the standard deviation.

In Figure 6.3 we show the computed global score, *i.e.* `Pseudo_SUS` – it is a measure of the perceived usability. We applied a normalisation factor that would account for the presence of an extra question (*i.e.* at step 4 we multiplied by $\frac{10}{11}$).

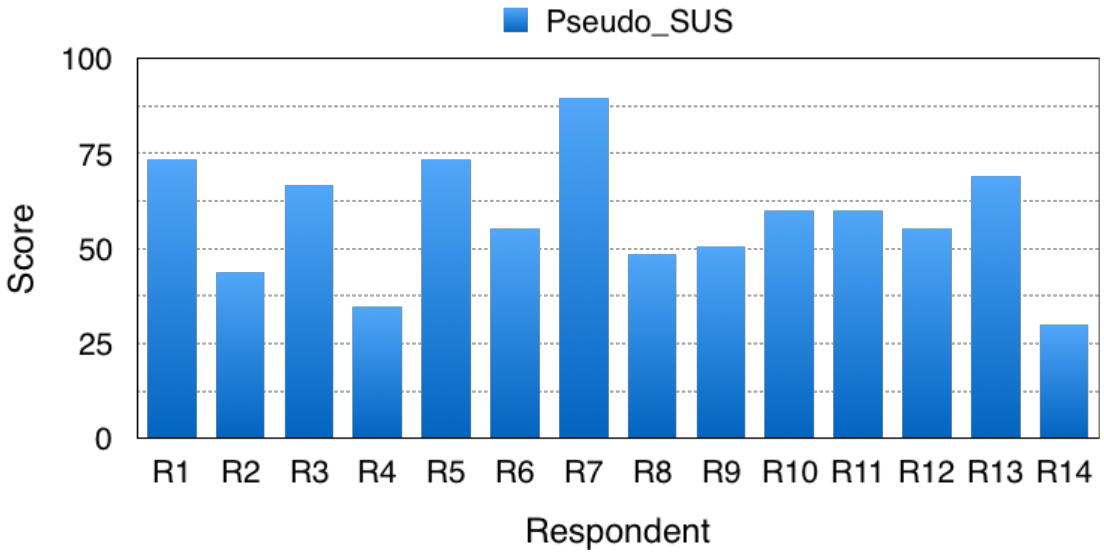


Figure 6.3: Showing the `Pseudo_SUS` score computed for each respondent (R1 – R14) – 9 out 14 (64%) experts consulted responded with a positive or neutral overall assessment.

A useful information can be obtained by observing the cumulative frequency and percentage of the `Pseudo_SUS` ranking in the categories (0-10), . . . , (90,100) in Figure 6.5. Similar ways to present SUS results proved very useful in combination with acceptability scores, adjectives and grades [Bangor et al., 2008; Brooke, 2013; Sauro, 2018].

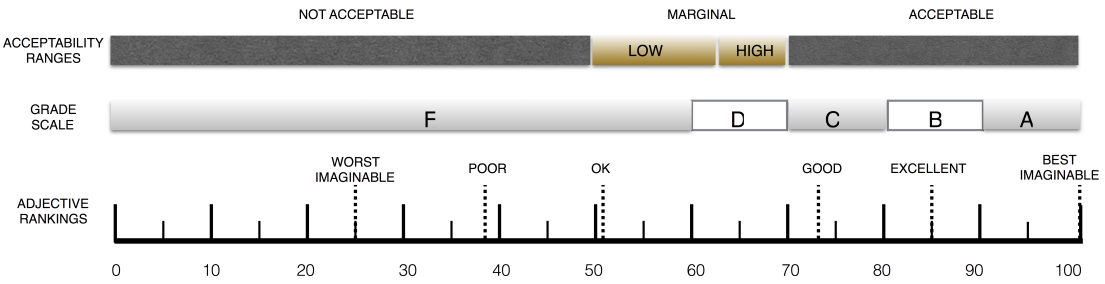


Figure 6.4: A comparison of mean SUS scores in relation with adjective ratings, grade ranking and acceptability scores

Source: [Bangor et al., 2008]

Figure 6.4 shows that by adding complementary information SUS results acquire more meaningfulness – the combination with other rating systems offers a reference to compare with and avoids the erroneous interpretation of SUS scores as percentages [Brooke, 2013].

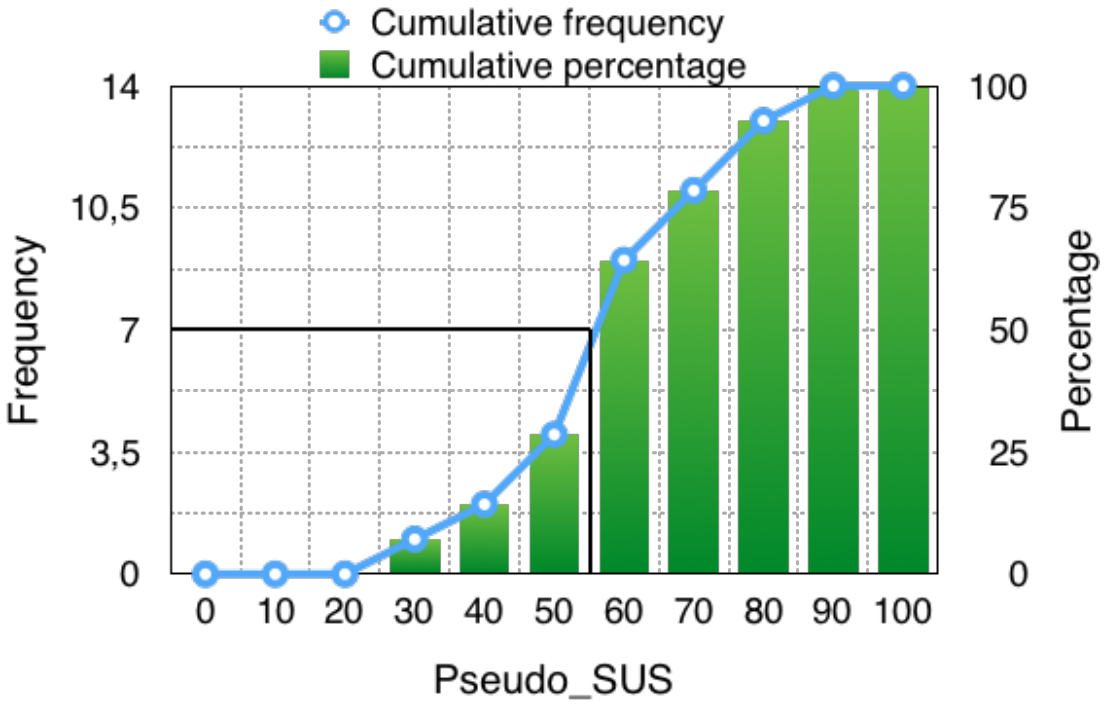


Figure 6.5: Showing the cumulative frequency and percentage of the `Pseudo_SUS` score ranking in the ranges (0-10), (10-20), ...,(90-100). The median value is also highlighted.

As showed in Figure 6.4 in an unaltered SUS an overall score above ∼51 would be considered fairly acceptable (*i.e.* OK) [Bangor et al., 2008]. Therefore, with $\texttt{Pseudo\_SUS}_{mean} = 57,5$ we could interpret our results quite positively. The score ranking in Fig. 6.5 confirms this. However, we are aware that the modifications applied to the questionnaire might affect the interpretation of our results. To mitigate that we presented the average score of the single questions in Figure 6.2. They show that the main features of our methodology applied in the context of EPOS are quite well understood. The consistency of the approach is recognised, although the application and the potential benefits of some key features (*e.g.* separation of concerns) are still debated. It should be noticed that the sample analysed is quite small and it might

include bias. However, these appear as reasonable results at this stage. Also, they provide us with clear indications about critical points for future assessments.

In the next sections we continue our evaluation by addressing the three dimensions individually: C, R and P.

6.2.2 Conceptual definition

The conceptual definition of the EPOS Canonical Core (CC) yielded an initial set of high-level concepts described in Chapter 5. We argued that such a set needs to evolve regularly in order to fulfil community requirements. To be included in the CC, concepts require agreed and harmonised definitions. The interaction processes that lead to such agreements are complex and time-consuming, thus determining the pace of the evolution. With the support of our methodology we aim to facilitate such processes and make them more manageable.

At present (October 2018) we record the first results and influences of such 'controlled' interactions. As the population of the EPOS CC progressed, and we show this in detail in Section 6.2.4, new requirements emerged together with issues that were not immediately manifest. By observing the content of the first CC users discovered capabilities and a new potential. For instance, they recognised they needed to discuss and agree the definitions of additional and more specific concepts. This was in some way expected as the initial set of Core Concepts was intentionally quite generic – it included high-level categories that should be further refined and specialised. As explained in Chapter 5, those specialisations would initially reside in the Boundary Regions and might become part of the CC whenever the participating communities recognise a need for including them.

We address such a process that enables 'promotion' of community concepts in the CC by observing a concrete application in Section 6.2.2.1. It shows the kind of dynamics that we envisaged and that should be observed and analysed over a longer period of time. A key benefit obtained by adopting the EPOS CC was the organisation of knowledge that could be derived from it. Core Concepts provide categories where domain-specific concepts can be hooked in. Therefore, domain knowledge (*e.g.* expressed and formalised in community vocabularies) can be combined and connected with the Core Concepts. A simple mechanism to link Core Concepts with

domain specific concepts is by labelling the first with keywords. For instance, in the population of the EPOS CC several keywords were supplied. In that phase there were no prior agreements and the keywords were mostly arbitrary and free-text based.

The adoption of a more structured process to choose and organise such terms, instead of free-text keywords, would provide considerable benefits. For instance, the employment of controlled vocabularies, a type of KOS (introduced in Chapter 3), would help with resolving synonyms.

In the next section we show how the application of our methodology to build and sustain the EPOS CC offered a way to steer and move forward the work initiated by the EPOS Vocabulary Task Force (VTF) that had been established to investigate the use of community vocabularies.

6.2.2.1 The EPOS Vocabulary Task Force

The VTF was initially set up, by gathering representatives (*e.g.* scientists and data experts) from each EPOS domain, in order to investigate the use of vocabularies in the corresponding communities. In a first phase its scope was quite broad and its focus not clearly defined. The activities of the VTF started with the collection of information about existing, standard vocabularies that could be of interest for the communities. However, it soon became clear that a more focused mission was needed.

By actively participating in the VTF we outlined the principles underpinning the construction of the EPOS CC and clarified how it supports the establishment of a shared vocabulary while taking into account existing community initiatives. We proposed to organise the activities of the task force by using our methodology. This triggered fruitful discussions that made the importance of the VTF evident as a framework to guide refinements and evolution of the EPOS Core Concepts and to share guidelines and best practices to organise knowledge.

Therefore, a plan was outlined and a number of concrete actions were identified. These are summarised below:

- Collection of existing vocabularies with a use-case driven approach – by restricting the focus to concepts and terms potentially related to the Core Concepts;

- Establishment of synergic harmonisation processes – that were previously carried out by independent thematic Harmonisation Groups (HGs);

- Definition of a mechanism to perform 'controlled' harmonisation – that is only where needed by requirements; and

- Identification of actors involved, their roles and responsibilities – to stimulate engagement and sustain commitment.

It was recognised that such a plan would require the establishment of a continuous process, where the commitment of key roles is a crucial factor for its success. By observing the dynamics of the VTF we concluded that such engagement could be sustained and supported by our approach leveraging a separation of concerns. For instance, a key scientist and a leader, remarked in one of the VTF meetings: "*I don't want to get involved in technical details about data models and such [...] but I believe that scientists should take the responsibility [of the content of the shared vocabulary]*". Such a feeling, widely shared among the participants, reflects the requirement for a focused approach that enables participants to address the concerns of their interest. In this case scientists clearly feel the need and the responsibility to lead the discussions about semantics and conceptual issues. At the same time they prefer to avoid being involved in the processes that formalise their conclusions into technical implementations. The latter is the realm of other experts.

Drawing on those requirements we proposed to apply our methodology in a process illustrated in Figure 6.6.

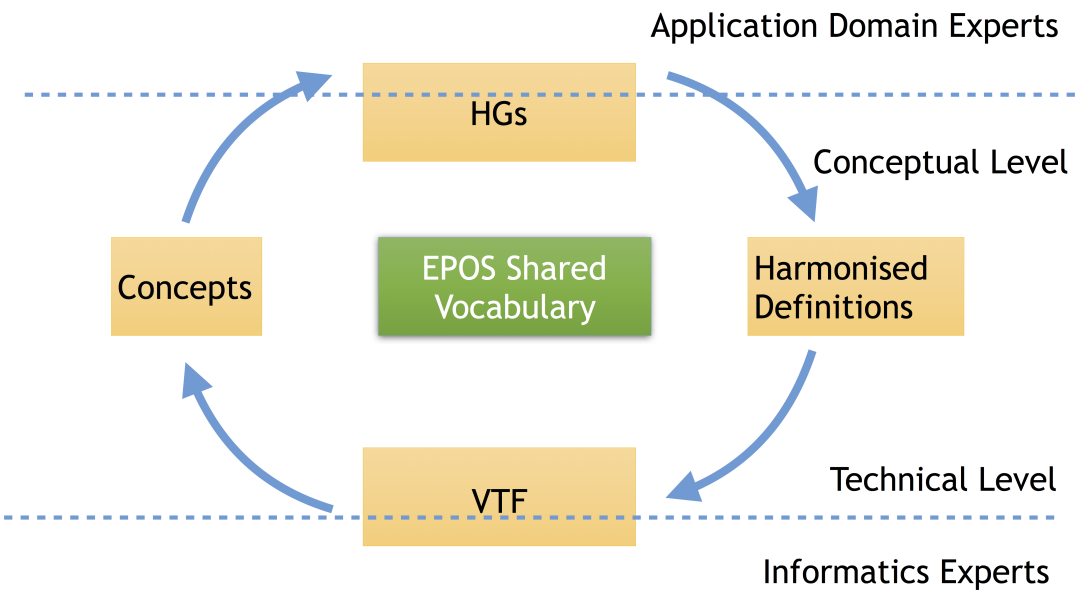


Figure 6.6: Sustaining shared vocabularies in EPOS.

In such a process the continuous interaction between the VTF and the Harmonisation Groups (HGs) is central and their roles are complementary. HGs are composed of domain experts, scientist and leaders and are in charge of the conceptual agreements. HGs had been initially constituted at an EPOS meeting[9] (March 2015) with the purpose to tackle harmonisation of EPOS assets. They were organised around 20 thematic areas with shared interests.

In October 2018 of the original 20, only 7 were still active, 6 in 'stand-by' and the remaining closed. Such a fragmentation is reflected also in their reported results. An analysis of the causes of such situation is out of scope in this research, however we want to highlight the difficulty to sustain the essential engagement of key leading roles in activities that are not in their primary focus. The approach illustrated in Fig. 6.6 with a close coordination and interaction with the VTF is an attempt to revitalise the activities of the HGs by providing them with clear focus and goals.

In such a process the VTF assumes a fundamental role and has the responsibility of the following tasks:

- Promoting and sharing methods and best practices to structure knowledge, *e.g.* by adopting standard vocabularies.

[9] https://www.epos-ip.org/events/epos-ip-project-tcs-ics-integration-workshop

- Evaluation of use cases and related concepts *e.g.* to decide whether the current set of Core Concepts have the required coverage.
- Periodical review of the set of Core Concepts and identification of potential modifications, *e.g.* to spot new candidates for promotion in or removal from the CC.
- Proposal for candidates to HGs, *e.g.* to enable their assessments about overlapping or related terms.
- Interaction with IT experts *e.g.* to formulate requirements to implement changes of the CC.

Similarly the tasks of the HGs, that represent the users and scientific community, are:

- Propose cross-disciplinary use cases, that are expected to be covered by the CC.
- Evaluate VTF's proposals for change, *e.g.* to assess whether actions are required.
- Harmonise definitions of selected concepts, *e.g.* when sufficient agreement is reached about new concepts to be included in the CC.
- Ratify changes of the CC, *i.e.* scientists and users have the final responsibility about the content.

In summary, the VTF assumes a role of mediator between technical implementation and conceptual definition. It helps identify areas that require harmonisation and trigger reconciliation processes that are performed at conceptual level by scientific experts. This allows them to keep the interactions focused and to avoid attrition of expertise.

This approach was proposed to VTF participants and then presented to a broader audience at the meeting mentioned in Section 6.2 (October 2018). A dedicated survey was conducted on this topic and collected responses from 10 participants (out of 35) – the low return rate might be attributed to the fact that this was the last of 4 surveys. The questionnaire and results are presented in Figure 6.7.

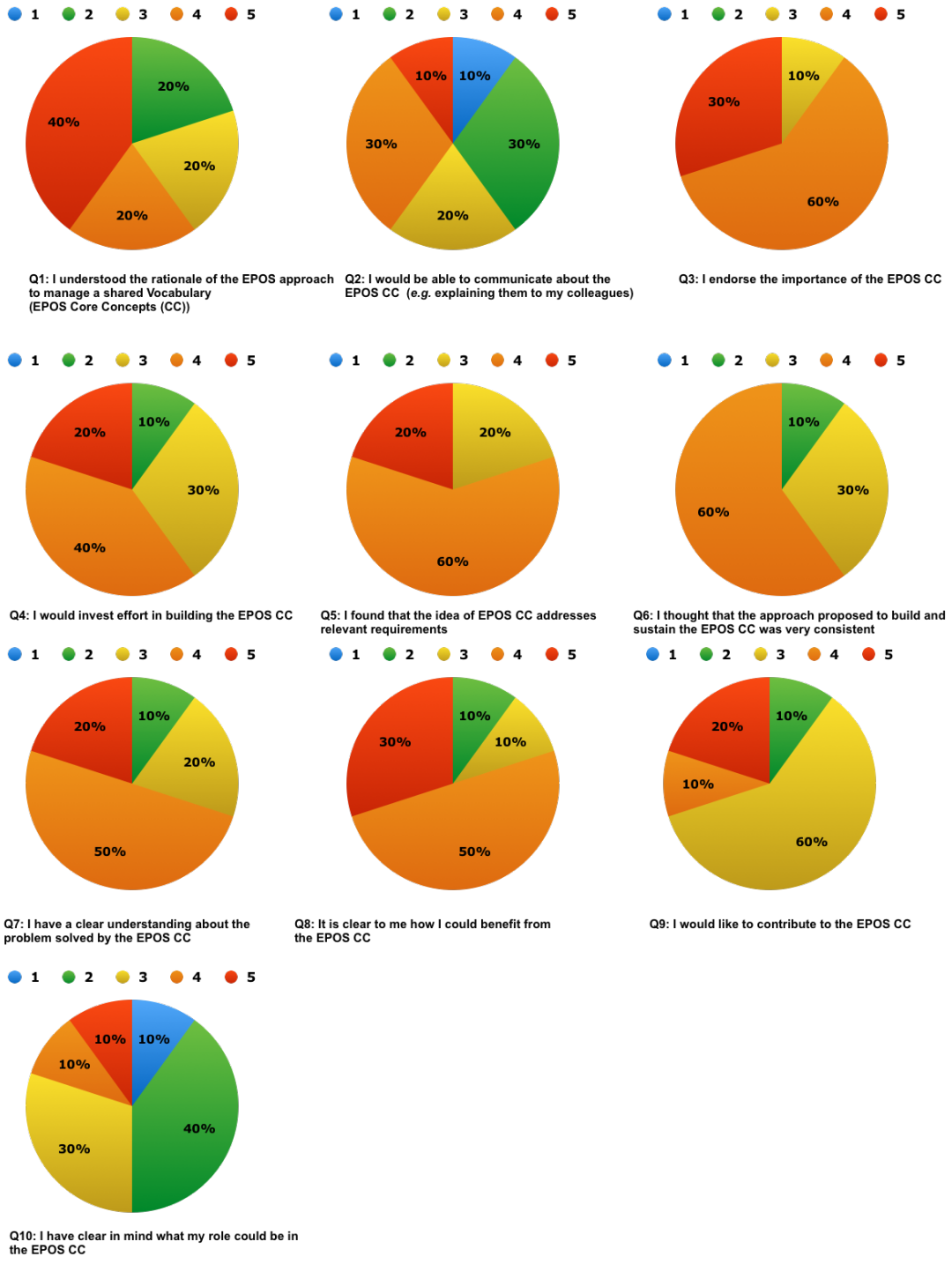


Figure 6.7: Introducing the questionnaire and the results of the survey: "*Evaluation of the EPOS Vocabulary approach*". The scale of results' values ranges from 'Strongly disagree' (1, blue) to 'Strongly agree' (5, red)

Figure 6.8 shows the average score obtained in the different questions.

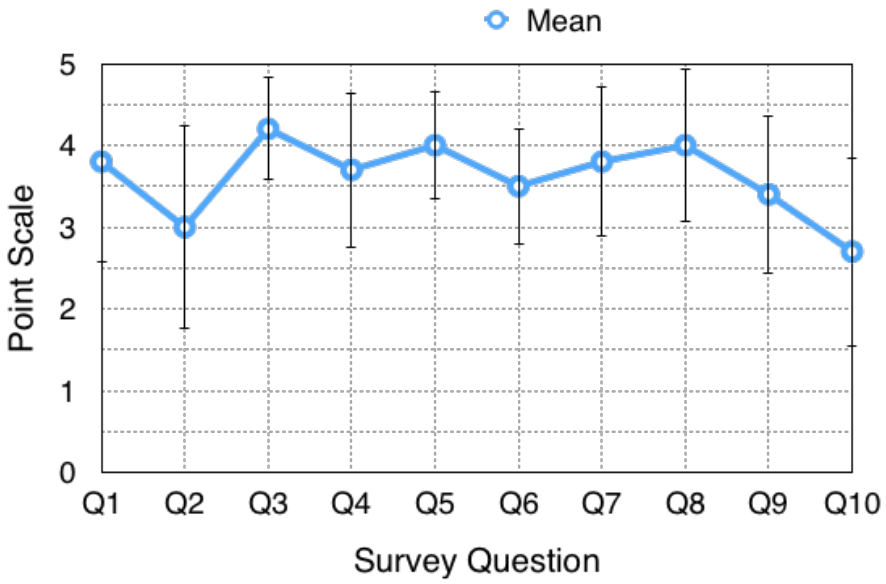


Figure 6.8: Showing the average (mean) points obtained in each question (Q1 – Q10) for 10 responses in every case. Error bars refer to the standard deviation.

Figure 6.9 illustrates the `Pseudo_SUS` score for each participant.

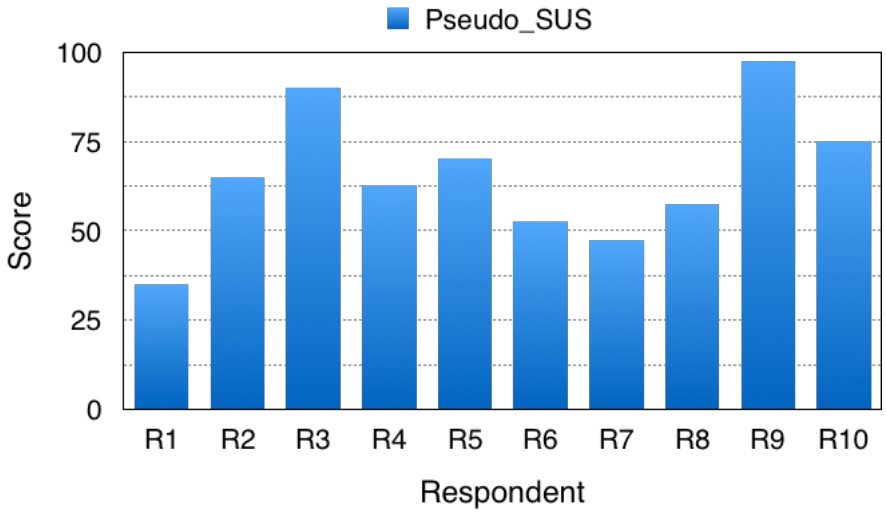


Figure 6.9: Showing the `Pseudo_SUS` score computed for each respondent (R1 – R10). 80% of the responses has an overall neutral or positive assessment.

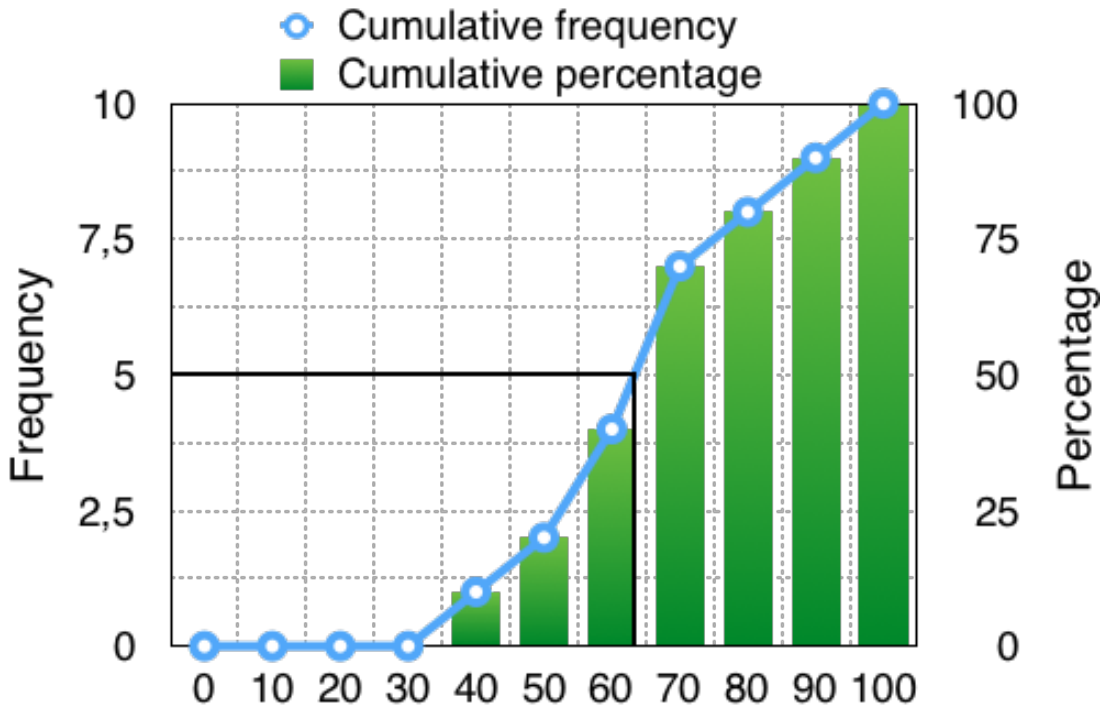


Figure 6.10: Showing the cumulative frequency and percentage of the `Pseudo_SUS` score ranking in the ranges (0-10), (10-20), ...,(90-100). The median value is also highlighted.

Figure 6.10 shows the cumulative frequency and percentage of the `Pseudo_SUS` score. The small sample size, the low return rate of participants and the potential presence of atypical respondents are important factors to take into account in this evaluation. Nevertheless, with a median score between 60 and 70 this is an encouraging preliminary result that can be exploited as a benchmark for future assessments. Those will be required to validate the approach and evaluate its efficacy to sustain the engagement of the participants. In the next section we move to the evaluation of the representation dimension.

6.2.3 Representation

The representation dimension was addressed with the establishment of EPOS-DCAT-AP. This model is widely adopted in EPOS and it enabled the active participation of several contributors who helped refining and tuning it in progressive stages. This suggests that our chosen strategy proved successful and promoted collaborations among

technical experts. Thanks to its features EPOS-DCAT-AP fits the requirements of broader communities beyond Earth sciences. For these reasons such a representation was presented and made available to a wider audience and it obtained positive considerations [Trani et al., 2018c,b].

Also, the requirements addressed by EPOS-DCAT-AP were recognised and taken into account in the revision of DCAT performed by the DXWG [W3C-DXWG, 2018]. For instance, the new DCAT revision accommodates key features of EPOS-DCAT-AP, *e.g.* support for additional resource types and data services [Beltran et al., 2018]. A discussion with the DXWG has been initiated [Browning, 2018].

These are preliminary evidence that demonstrate the influence and the validity of our approach to tackle representation. To evaluate the usability of EPOS-DCAT-AP we conducted a survey that received responses from 20 participants (out of 35) and whose results are presented in Figure 6.11. It is interesting to analyse them in the light of the preliminary assessments presented in Chapter 5.

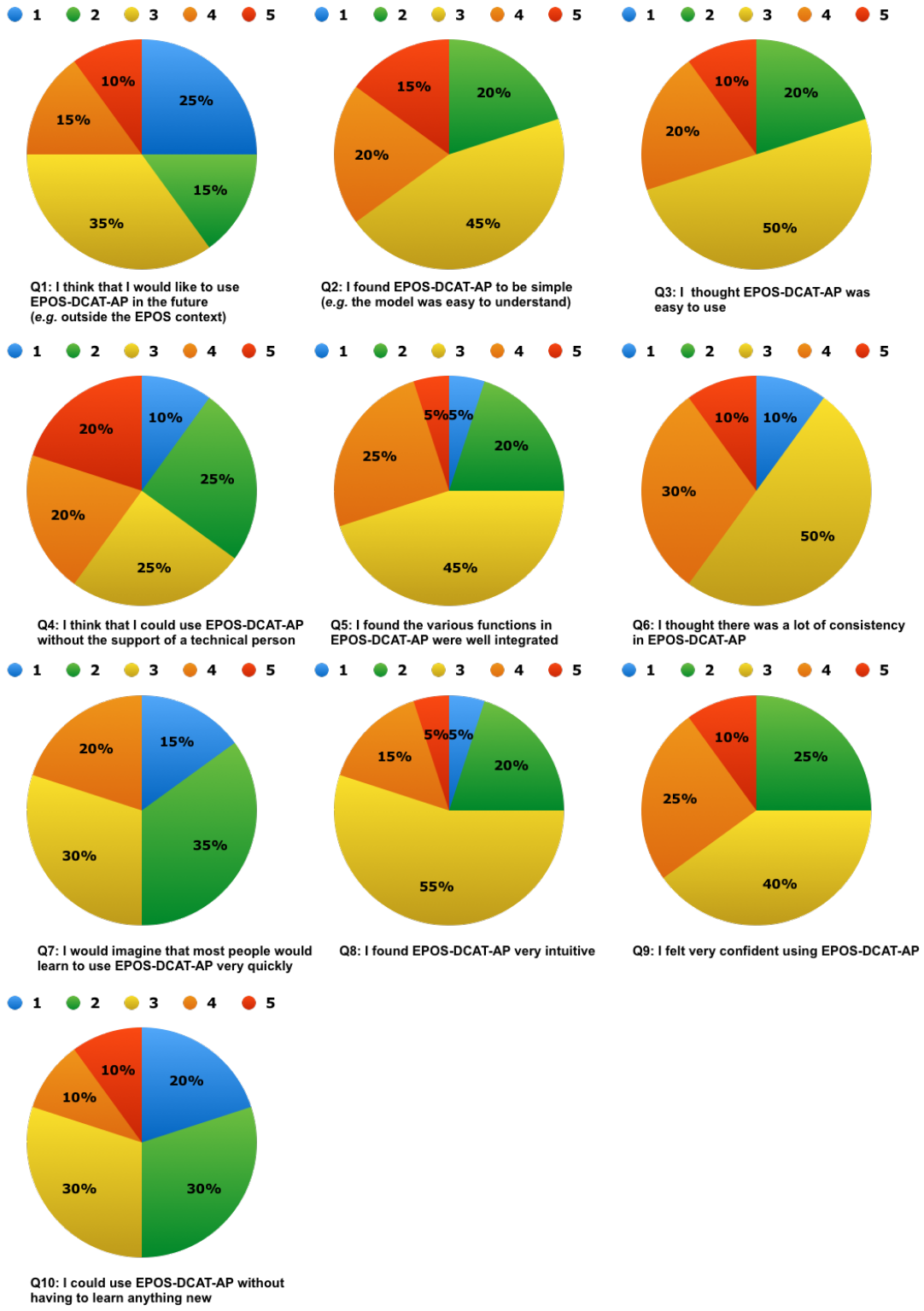


Figure 6.11: Introducing the questionnaire and the results of the survey: "*Evaluation of the EPOS-DCAT-AP*". The scale of results' values ranges from 'Strongly disagree' (1, blue) to 'Strongly agree' (5, red)

Figure 6.12 shows the average score in each question.

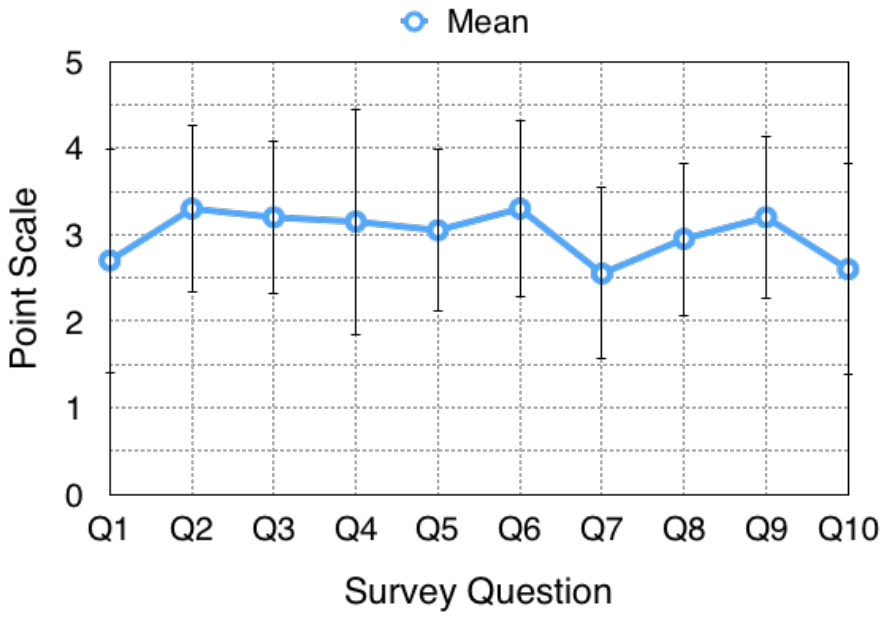


Figure 6.12: Showing the average (mean) points obtained in each question (Q1 – Q10) for 20 responses in every case. Error bars refer to the standard deviation.

Figure 6.13 presents the `Pseudo_SUS` score obtained by each respondent.

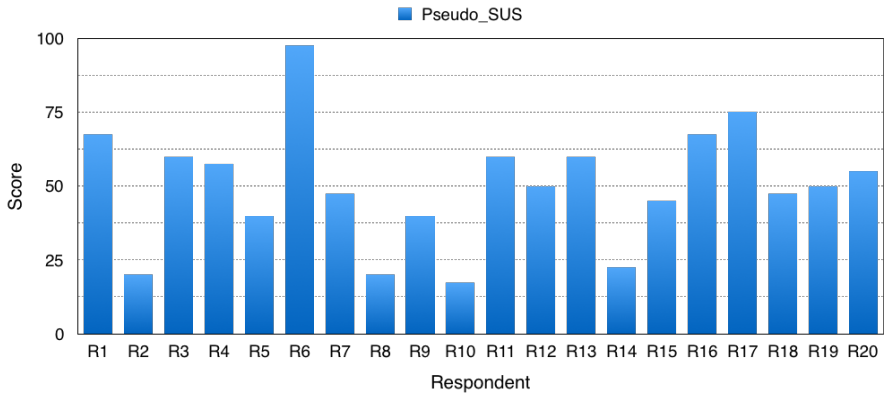


Figure 6.13: Showing the `Pseudo_SUS` score computed per each respondent (R1 – R20). 45% of the responses has an overall neutral or positive assessment.

In Fig. 6.14 we show the cumulative frequency and percentage of `Pseudo_SUS`.

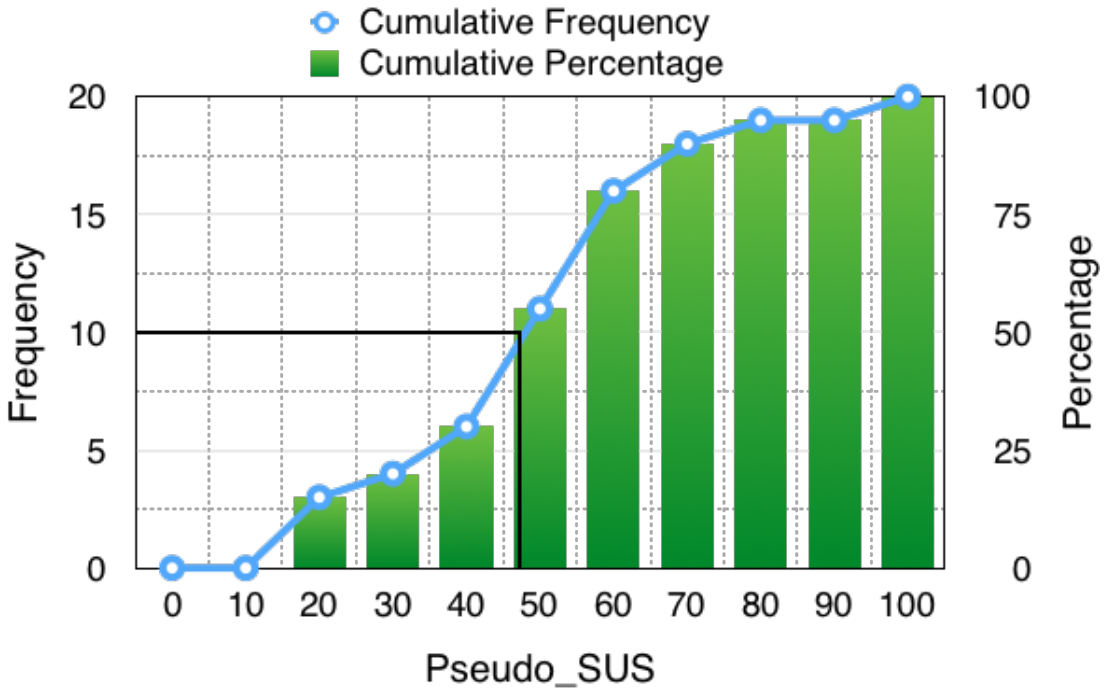


Figure 6.14: Showing the cumulative frequency and percentage of the `Pseudo_SUS` score ranking in the ranges (0-10), (10-20), . . . ,(90-100). The median value is also highlighted.

The results show that improvements are still required. In particular, the limited documentation might affect significantly the perceived usability. A first draft of a comprehensive specification and documentation is now available (January 2019) [Paciello et al., 2019].

6.2.4 Population

As described in Chapter 5 the population process has been organised and carried out in incremental stages. After an initial phase where the main goal was to assimilate and test procedures, successive steps followed that helped refining requirements, tuning and validating the approach to tackle them. For instance, the requirement for automated tools became evident in order to support harvesting operational large-scale populations. Tools were introduced to enable metadata validation (*i.e.* based on SHACL) and to aid the construction of metadata documents (*e.g.* Web Metadata Editor[10]). Once produced

[10] `http://epos.cineca.it/apache/mde/public/index.php`

and validated such metadata documents were collected in a collaborative environment, *i.e.* GitHub, in order to be curated and then ingested into the EPOS metadata catalogue. In this pre-operational phase the manual curation proved essential to guarantee the quality of the metadata content. It helped identify inconsistencies and highlighted issues that were then fixed in collaboration with the providing communities. We foresee that the role of curators will be alleviated by automation and accurate tools. Nevertheless, it will remain important also in the operational phase where procedures will be to a large extent automated.

The official access channel of EPOS is its portal[11] that builds on the EPOS metadata catalogue. However, the availability of the metadata in RDF format allows us to easily import them in a compliant data store. For instance, in Fig. 6.15 we illustrate an example of population with a Neo4J database (a popular graph database introduced in Chapter 3) equipped with a plugin[12] that enables Semantic Web capabilities. Thanks to its supported visualisation tools it provides users with advanced features for presentation of content, interactive query and browsing of results. In this way it might assist (meta)data curators in assessing the quality of the imported populations. The results shown in Figure 6.15 are obtained by submitting the query in Listing 6.1 expressed in the Cypher query language [Neo4J, 2016].

Listing 6.1: Cypher query that produces the graph in 6.15

```
MATCH (n:`http://www.w3.org/ns/dcat#Dataset')-[:`http://www.w3.org/ns/dcat#distribution']->
(d:`http://www.w3.org/ns/dcat#Distribution'),
(n:`http://www.w3.org/ns/dcat#Dataset')-[:`http://www.w3.org/ns/dcat#theme']->
(c:`http://www.w3.org/2004/02/skos/core#Concept')
where c.`http://www.w3.org/2004/02/skos/core#prefLabel'
contains "Magnetic"or c.`http://www.w3.org/2004/02/skos/core#prefLabel'
contains "Seismic" or c.`http://www.w3.org/2004/02/skos/core#prefLabel'
contains "GNSS" and date(LEFT(n.`http://purl.org/dc/terms/created',10)) < date("2018-11-19")
return n,c,d
```

[11] https://epos-ics-c-beta.brgm.fr/

[12] https://github.com/jbarrasa/neosemantics

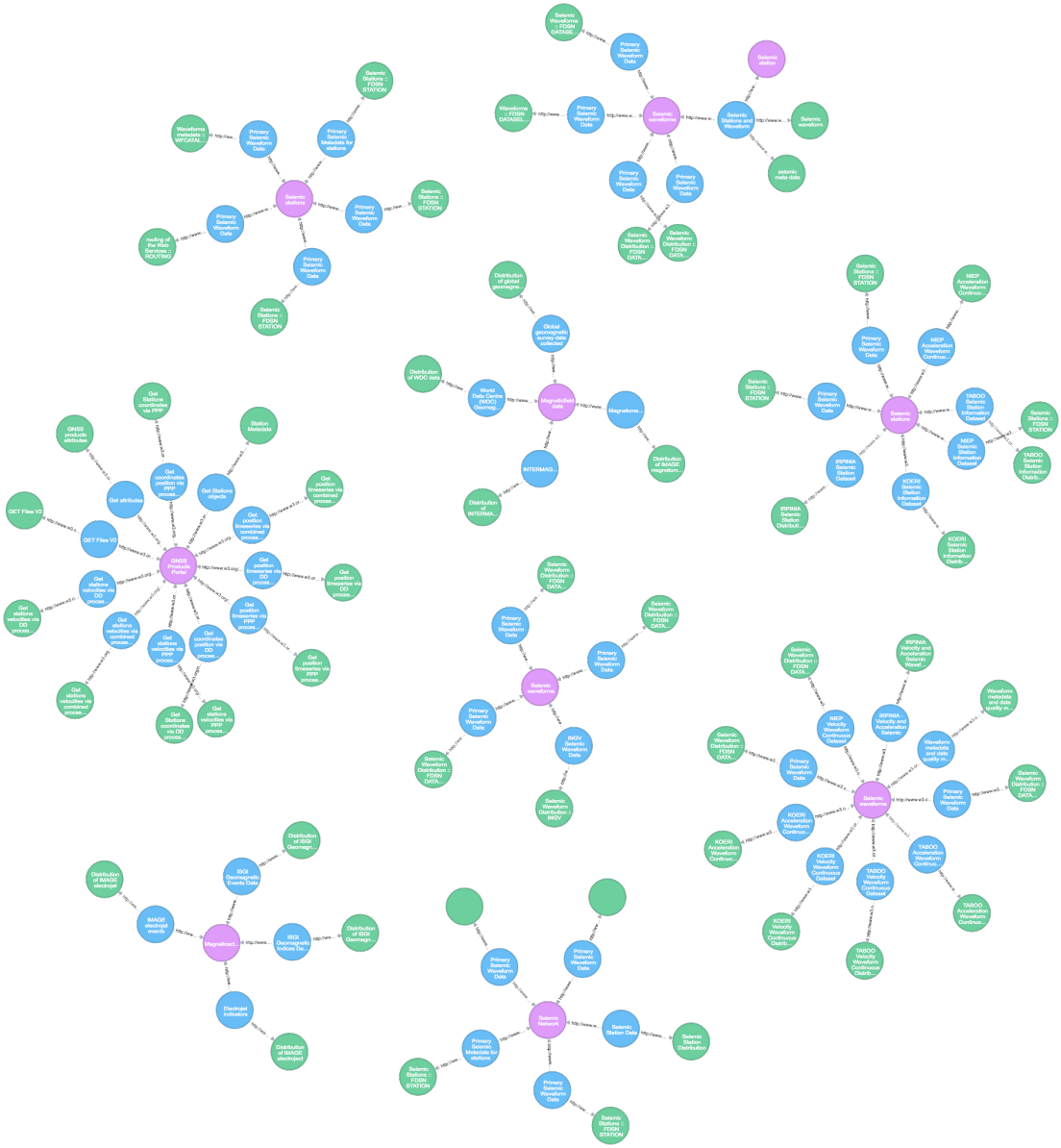


Figure 6.15: Showing the results of a query that matches Datasets (blue) and related Distributions (green) created before a specified data (*i.e.* 19-Nov-2018) whose associated Concepts (purple) contain the terms “Magnetic”, “Seismic” or “GNSS”

Table 6.2 summarises the status of the current (October 2018) population of the main entities. Dimensions are still modest compared with the expected final scale – the current graph contains ∼3200 nodes and ∼4100 relationships. Inevitably, those are going to increase in volumes and variety of types starting from the next population phase that will include additional entities and relationships (*e.g.* Software, Publication, Facility, Equipment). As in the previous population phases a restricted number of communities have been selected to start the process and test the ingestion of the new entities, others will follow promptly.

Entity name	**Population**	
	March 2018	October 2018
Person	86	99
Organisation	32	38
WebService	74	97
Dataset	NA	122
Operation	NA	144
Distribution	NA	153

Table 6.2: Population statistics (October 2018). Relative to the data reported in Chapter 5 (March 2018) we notice an increase in volumes and the presence of additional entities, such as Dataset, Operation and Distribution.

Similarly to the other dimensions, we present the results of the usability survey on population and in particular addressing the tools that enable such processes. It collected responses from 13 participants (out of 35). Figure 6.16 illustrates the questionnaire and the response statistics.

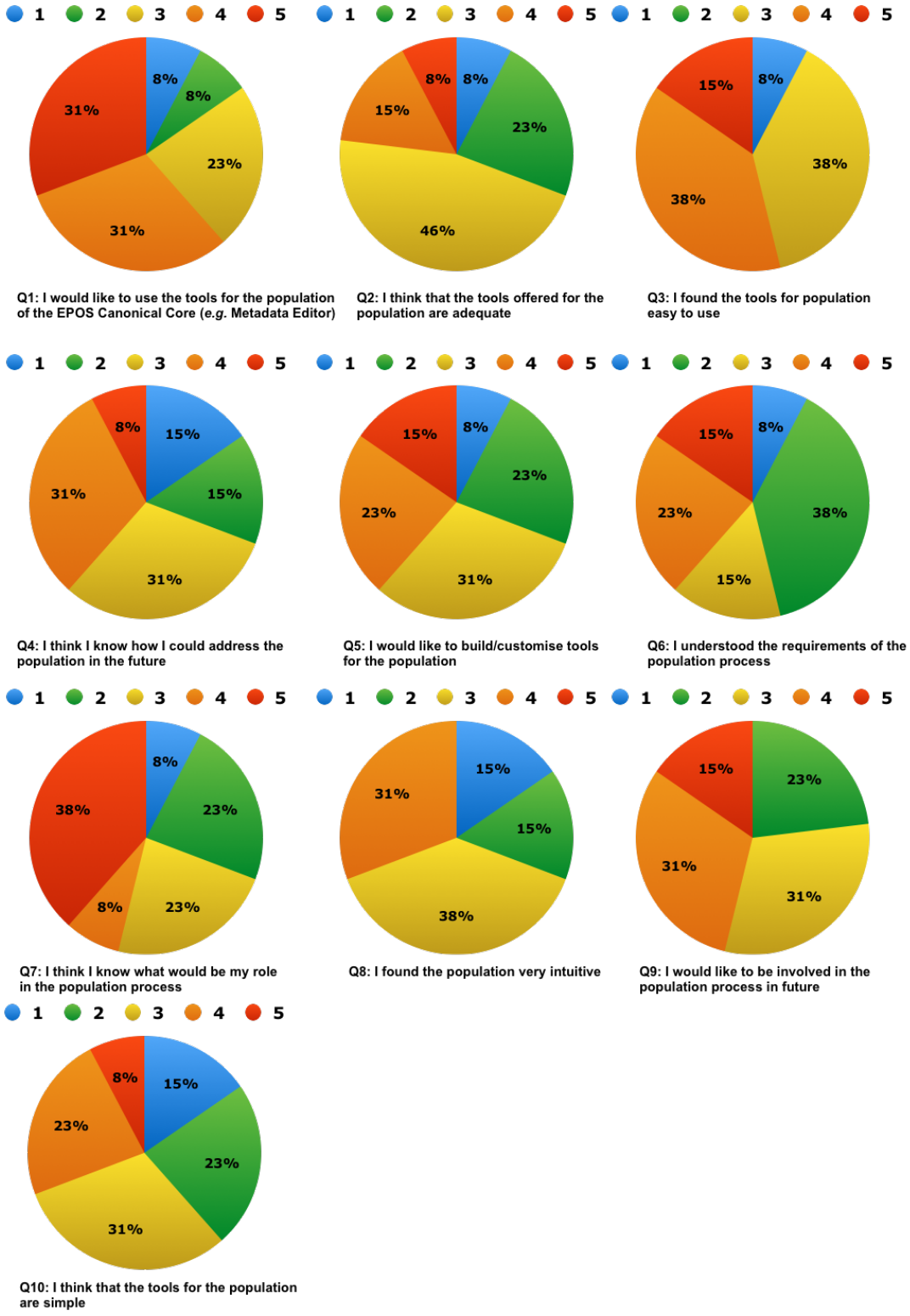


Figure 6.16: Introducing the questionnaire and the results of the survey: "*Evaluation of tools for population of EPOS Canonical Core*". The scale of results' values ranges from 'Strongly disagree' (1, blue) to 'Strongly agree' (5, red)

Figure 6.17 shows the average score in each question.

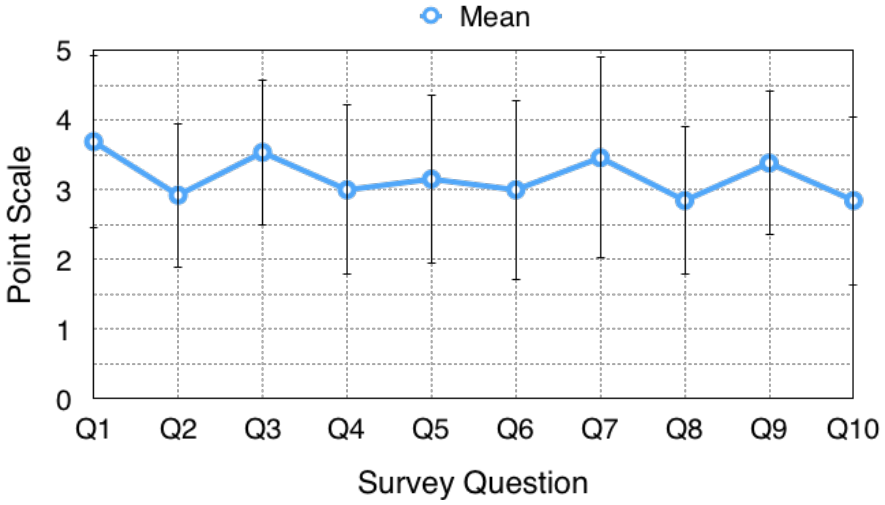


Figure 6.17: Showing the average (mean) points obtained in each question (Q1 – Q10) for 13 responses in every case. Error bars refer to the standard deviation.

Figure 6.18 presents the `Pseudo_SUS` score obtained by each respondent.

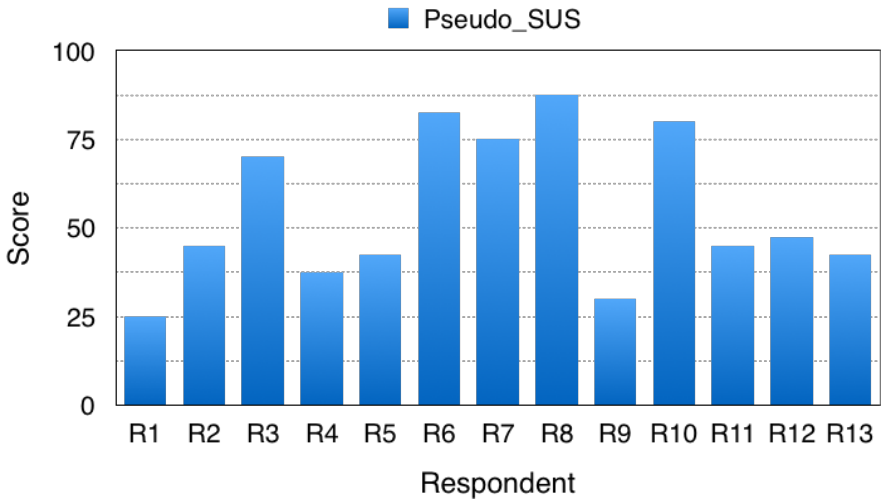


Figure 6.18: Showing the `Pseudo_SUS` score computed for each respondent (R1 – R20). 38% of the responses express a positive assessment.

In Fig. 6.19 we show the cumulative frequency and percentage of `Pseudo_SUS`.

The results obtained show that there is room for substantial improvements in the population dimension. Most likely, these results reflect the fact that the population

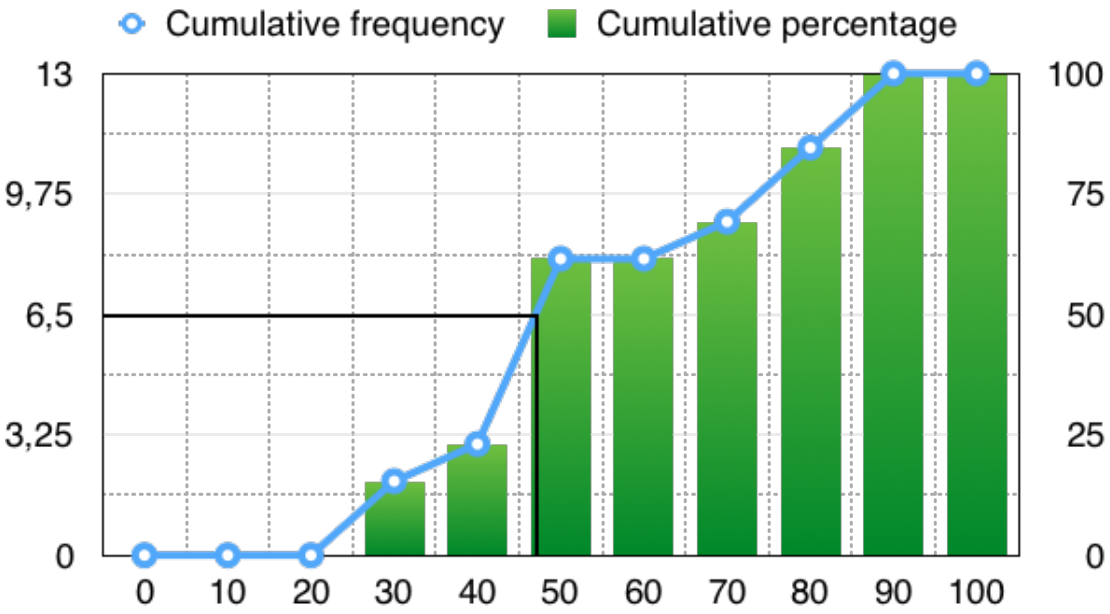


Figure 6.19: Showing the cumulative frequency and percentage of the `Pseudo_SUS` score ranking in the ranges (0-10), (10-20), ...,(90-100). The median value is also highlighted.

is still too much of a manual process. Assessments will be required when the set of automated tools will be enriched and the population streamlined.

6.3 Conclusions

In this chapter we provided an assessment of our methodology by addressing several aspects. In this first part we reported the progress and results achieved by an established focused seismological core (*i.e.* `WFCatalog`). They show that the requirement for the definition and usage of a set of agreed concepts that enable collaboration is valid and its utility recognised. We then moved to a wider context that introduced additional challenges. That is the main focus of this research and we explained the difficulties to perform a comprehensive evaluation at this stage. Those are related to the low maturity of our methodology (currently in its early days) and the scope of application that we could achieve in the limited frame of this research. For these reasons we proposed to set up an evaluation framework that can be harnessed in future assessments that would require repeated and regular evaluations.

From the data and experience collected so far we can draw some conclusions and highlight aspects and critical points that will deserve attention. The rationale of our strategy to encourage collaboration starting from a set of Core Concepts proved successful and obtained a good uptake. Although we are just at the early stages and our final goal targets the long-term, we can already identify the kind of fruitful dynamics that we had envisaged. For instance, the partitioning of the problem space into three dimensions offered a way to convey focused effort from the interested actors. They recognised the benefits of this approach and are willing to continue adopting it. We reported about discussions initiated, challenges addressed and issues solved thanks to our methodology. This initial phase shows that we were able to engage the communities and the stakeholders with the right arguments and motivations. To monitor their interest and engagement over the long period we propose to exploit surveys together with more 'observational' techniques (*e.g.* Critical Incident Technique). However, to increase the robustness and validity of the results broader and varied samples should be targeted. For instance, an important test for assessing the effectiveness of our approach would be to transpose it to different communities with heterogeneous levels of maturity and interaction.

In the next chapter we outline our future strategy – a central part of it is related to the evaluation and refinement processes that will enable us to maintain alignment with communities and their requirements.

Chapter 7

Conclusions and Future Work

In this thesis we have discussed extensively the importance of information and knowledge sharing collaborations and their key role for science and research. We highlighted how the human component is essential in the interaction processes that lead to pooling of knowledge and expertise. Similarly, we showed how automated methods and structured approaches are valuable supports to sustain those processes. Such views were corroborated by the experience acquired and by the lessons learned pursuing this research. In this chapter we draw together key findings and arguments that supported our initial thesis. We summarise implications and impact of our approach on contemporary research and we then conclude by outlining future directions and possible developments with a long-term vision.

7.1 Achievements and influences

In the following sections we restate key achievements of this research and emphasise their influences. The investigations performed in this thesis, in particular in Chapters 2 and 4, reveal that the requirement for a structured and professional approach to sustain scientific collaboration is stringent. In Chapter 2 we reported several initiatives that pursue such a goal by adopting diverse strategies. Those are often tailored to specific needs and therefore difficult to scale and apply in broader contexts. At the same time a wealth of tools provides a strong technological support for various elements as we showed in Chapter 3. Such an abundance might be a trigger that incentivises researchers to tackle the collaborative challenges from a technical perspective. For

instance, this was the initial VREs' approach. However, such an approach offers a limited vision of the issues involved. In Chapter 4 we described how we met the socio-technical challenges by tackling the requirements of a seismology community.

We motivated the value of an *alignment of technical and intellectual efforts* – a successful approach to achieve collaborations must consider and target both those aspects, thus aiming for a well-balanced mix where technology does not hinder but rather serves, stimulates and facilitates human interaction. To achieve this *we proposed a methodology*, described in Chapter 5, that builds on a solid conceptual ground derived from the CSCW research. Our methodology empowers researchers to build and maintain a shared stable context or Common Information Space, which can be extended and linked with domain-specific knowledge embedded in boundary objects and knowledge artifacts. It *combines recognised intellectual capital with mature methods and related technologies* such as the Semantic Web and Linked Data.

By integrating organisational, conceptual and technical aspects in a consistent framework that recognises the centrality of the human interactions we demonstrated preliminary but substantial benefits. We reported adoption and uptake in contexts of different maturity and progressive complexity. They resulted in a good engagement and commitment obtained in a relative limited timeframe. In Chapter 6 we assessed those achievements and outlined a strategy to perform future evaluations. Those will be required to assess the capability of our methodology to support sustained collaboration over a longer period.

The applicability of our approach in different settings with consequent improvement in cross-collaboration derived by sharing common methods delivers great added value compared to one-off, tailored solutions. Refinements and improvements will be needed, for instance to sharpen and detail processes and to support them with more automation. Those will be natural extensions that, thanks to the flexibility and modularity of our approach, could be integrated incrementally without disrupting the methodology's philosophy and conceptual backbone.

7.1.1 Aligning socio-technical challenges

The first important conclusion that we drew after the initial investigations and experience acquired in a seismological context was the recognition of the value and impact

of socio-technical challenges. As we reported in Chapter 4 our first attempt to establish collaboration by pooling knowledge in a well-organised and focused environment, *i.e.* the ORFEUS-EIDA federation, exposed critical non-technical issues. To achieve engagement and sustain effort the organisational context must be taken into account – it is a catalysing element to stimulate agreement and an essential support to align priorities.

Those findings deeply shaped our strategy and influenced the direction of our research. They made us focus on *a conceptual framework that would support agile processes and layered Governance with distributed responsibilities*. To achieve this we advocated the establishment of a Canonical Core as a foundation for collaborative information and knowledge sharing connected with a set of flexible and dynamic Boundary Regions. Participants in the collaboration are in control and hold the responsibility about the content of the Canonical Core. At the same time they seek agreements on the interfaces with the Boundary Regions. Governance supports the necessary underpinning organisational processes by identifying current issues and by harnessing experts' help to tackle them.

7.1.2 Seismological waveform FAIRness

Another important achievement obtained in the initial phases of this research was the establishment of *a canonical representation* of seismic waveform features, *i.e.* `WFCatalog` – it was our first attempt to form a shared core with agreed concepts and definitions. `WFCatalog` offers users a canonical representation of seismological waveforms that supports efficient and streamlined data discovery and access. Albeit of limited scope, that representation yielded good results and enabled improvements and advances. We described `WFCatalog` in Chapter 4 and in Chapter 6 we reported evidence of success of our approach that was assimilated and adopted as an operational service across Europe and led to global consideration of adoption.

The outreach obtained and the data collected in a short operational timeframe are convincingly positive [Schuh et al., 2018; Vecsey, 2018; Atkinson et al., 2018]. For instance, we discussed how `WFCatalog` enables FAIRness of seismological waveform data [Trani et al., 2017; Koymans et al., 2018]. That path will be continued and exploited further in H2020 projects such as EOSC-hub [Trani and the EPOS-ORFEUS-

CC Team, 2018] and ENVRI-FAIR[1]. These results encourage us to believe that further potential benefits will be experienced in the near future when uptake by the user communities and usage of the service will both increase as a result of `WFCatalog` being incorporated in methods. New developments building on `WFCatalog` have been planned and the sustained commitment of the EPOS-ORFEUS community will facilitate them.

7.1.3 Focused interactions

Our research showed that key features that enable effective collaboration are focused interactions. We proposed that these are driven *by focusing on developing conceptual agreements*. In particular, they are based on a set of agreed Core Concepts sufficiently described to cover the principal use cases, but abstract enough to avoid locking discussions into unnecessary detail. Local extensions and specialisations connected with Core Concepts offer precision required by domain-experts – they enable different viewpoints and leave space for experiment and innovation from within a coherent framework.

We started by pioneering a methodology, which is summarised below, in a seismological context and demonstrated its utility. We reported commitment and operational adoption that will continue to be supported by an established organisational framework. Similarly, we presented benefits generated in the early stages of adoption in the context of EPOS. Also in that context our approach was positively evaluated and will continue in next phases of that multi-disciplinary research infrastructure.

Establishing and maintaining a valuable asset, such as the Canonical Core (CC), is a demanding intellectual challenge, as explained in Chapter 5 [Trani et al., 2018a]. We demonstrated that the benefits generated are worth the investment. *Not only the CC is functional to and instrumental for an improved communication among participants of a collaboration, but its construction process is extremely valuable. It stimulates thinking and encourages communities to make implicit knowledge emerge.* For instance, we reported how in EPOS this triggered discussions about definition of a common vocabulary and agreements about shared terminology.

We offered a methodology to build, manage and maintain the CC that decouples its

[1] `envri-fair.eu`

three components: Conceptual definition, Representation and Population. We applied such a partitioning throughout this thesis and showed its value for analysis and design – it helped us reviewing literature and it was a key principle in the engineering of our solutions. *Drawing the attention of the users to selected viewpoints based on their interests established a clear focus that helped them overcome impasses. It helped us retain their engagement and made issues more manageable.* Experience and evidence acquired suggest that we will continue that approach in the future *e.g.* in the EPOS context and beyond [Atkinson et al., 2018; Magagna et al., 2018].

7.1.4 Representing agile agreements

Once we had defined our methodology to build a Canonical Core we designed, applied and evaluated a solution to represent its content *i.e.* EPOS-DCAT-AP. In Chapter 5 we described the properties that such a representation should expose in order to balance between consistency and agility. Those two characteristics are required in order to *support stable interactions and at the same time not inhibit innovation.*

The combination of RDF with SHACL proved a valuable and successful solution. The first offers mechanisms to represent open knowledge whereas the latter provides flexible constraints to shape agreements without breaking the philosophy of an evolutionary approach. SHACL can be used to validate data graphs according to specified requirements, moreover it has advanced features that can be harnessed to define and optimise functions that enable knowledge filtering, *e.g.* according to pre-defined (SPARQL) patterns [Rashid et al., 2018; Knublauch et al., 2017]. An active community and interesting new developments with the endorsement of big technology players, *e.g.* the integration of GraphQL and SHACL, will facilitate continued adoption and support [Facebook Open Source, 2018; TopQuadrant, 2018].

Another principle applied in the design of our representation was the *reuse of existing conceptual bundles* and this yielded an extension of a standard vocabulary *i.e.* DCAT by including and leveraging wide-spread vocabularies such as Schema.org. EPOS-DCAT-AP was successfully applied in the context of a large research infrastructure. And the positive results and outreach obtained combined with a dialog and contributions established with broader communities of experts [Trani et al., 2018c,b; W3C-DXWG, 2018] suggest that it will be continued in EPOS and beyond by pro-

moting compatibility with broader standards [Quimbert et al., 2018; Atkinson et al., 2018]

7.1.5 Summary

To conclude our summary we can assert that the main goals of this research have been met. We defined, applied and assessed a *methodology* to support scientific users, data experts, designers and technical architects in their collaborative work. We empowered researchers who want to collaborate across discipline boundaries by offering *methods to build and maintain Common Information Spaces*. We targeted them with *personalised views* that promote their *expert engagement* by offering them *control and responsibilities* where needed, thus creating a *sense of ownership*. We helped them by lowering barriers to building and sustaining cross-disciplinary collaborations and by making challenges manageable. We pioneered the methodology to move from *conceptual agreements* to *technical implementations* and offered relevant tools to support the process. Our approach has been applied and evaluated in different contexts, however, more work remains to be done and in the next section we offer our vision for its continuation.

7.2 Future outlook

Our research has developed a foundation for improving computer-supported collaboration across discipline, organisational, and national boundaries. It already engages several organisations developing particular collaborations. We showed how several aspects of this foundation have been already adopted by targeted communities who will continue to depend on those components and motivate, if not resource, their sustainability. Such an endorsement will be an opportunity to improve aspects of our approach and to perform wider and more complete assessments. We believe that as a result of our research a novel collaborative culture can be initiated where an improved awareness of a shared information and a shared terminology will be fundamental elements.

7.2.1 Information-Powered Collaborations

By progressing towards innovation and collaboration driven by conceptual agreements, we foster the systematic establishment of Information-Powered Collaborations. Those will be driven not anymore by *ad hoc*, yet creative, efforts but rather by an improved and systematic approach to issues and challenges. IPC stakeholders and designated communities will be empowered by effective tools to tackle those and harness the power of the pooling of knowledge and resources. We envisage that such tools will be built incrementally with an adaptive rather than disruptive approach. Thanks to their active engagement users will be in control of migration and/or adaptation of existing methods in order to exploit the increased potential of the shared contexts.

Researchers will be able to annotate their findings and enrich them with links to the related Core Concepts. This will offer a mechanism to hook novel discoveries with their originating conceptual sources. Enhanced provenance services will enable explorations of multiple viewpoints of the resulting shared knowledge.

Similarly, participants of IPC will adopt tools to contribute their domain-knowledge to the CC and actively shape its content. The relationships with recognised Core Concepts will improve scientists' and practitioners' ability to communicate across boundaries. Finally, the extended experience acquired by users of IPC in their new collaborative practices will yield best practices and help formalise them into processes and methods, thus improving the resulting interactions.

7.2.2 Enhancing human processes

Whilst the primary role of human judgement will remain central, we envisage continuous improvements *e.g.* of the decision-support aids. The availability of a common knowledge base represented with recognised formalisms will facilitate the uptake and developments of tools of increased effectiveness. The technology advances will enable researchers to extract and filter desired information with improved performance.

Similarly, the agreement-forming processes will be empowered and facilitated by automated tools that will generate possible scenarios exploiting the available shared knowledge. When the data available reaches critical mass such tools, might exploit for instance Artificial Intelligence and Machine Learning to highlight critical paths. They might offer suggestions and spot new potential candidates for promotion in the CC,

e.g. by observing and analysing applications of concepts in scientific methods. In this way the evolution of the CC will be supported by automation. Nevertheless, the final choices and ownership of the conceptual space, CIS, underpinning the collaborations will remain in the hands of their involved stakeholders (*e.g.* scientists, practitioners, research infrastructure managers).

7.3 Conclusions

This work was stimulated and triggered by pressing needs and requirements that emerged by observing real working practices in modern scientific contexts. Our motivation was to improve the collaborative work experience and make it more effective and sustainable. We believe we demonstrated a methodology to achieve such improvements and paved the way for new collaborative culture. The evidence collected so far motivates us to continue investing in this approach.

Information-Powered Collaborations will keep flourishing driven by scientific challenges and societal demands, we now have a better understanding of their socio-technical dynamics and a *modus operandi* to unleash their potential.

List of Acronyms

BR Boundary Regions

CC Canonical Core

CCSDS Consultative Committee for Space Data Systems

CERIF Common European Research Information Format

CIT Critical Incident Technique

CKAN Comprehensive Knowledge Archive Network

CSCW Computer Supported Cooperative Work

DAMC Data Analysis and Metadata Computation module

DBMS DataBase Management System

DC Dublin Core

DCAT Data Catalog Vocabulary

DCMES Dublin Core Metadata Element Set

DCMI Dublin Core Metadata Initiative

DDSS Data Data products Service Software

DL Digital Library

DLMS Digital Library Management System

DLS Digital Library System

DOI Digital Object Identifier

ECMWF European Centre for Medium-Range Weather Forecasts

EIDA European Integrated Data Archive

EPOS European Plate Observing System

ERIC European Research Infrastructure Consortium

ESA European Space Agency

ESC European Seismological Commission

FAIR Findable Accessible Interoperable Reusable

FDSN International Federation of Digital Seismograph Networks

GEO Group on Earth Observations

GEOSS Global Earth Observation System of Systems

GNSS Global Navigation Satellite System

GPS Global Positioning System

HG Harmonisation Group

HPC High Performance Computing

HTC High Throughput Computing

IASPEI International Association of Seismology and Physics of the Earth's Interior

ICS Integrated Core Services

InSAR Interferometric Synthetic Aperture Radar

INSPIRE Infrastructure for Spatial Information in Europe

IPC Information Powered Collaborations

IRI Internationalized Resource Identifier

ISO International Organization for Standardization

IVOA International Virtual Observatory Alliance

KOS Knowledge Organization System

LD Linked Data

LDP Linked Data Platform

LOD Linked Open Data

MARC MAchine-Readable Cataloging

MUSTANG Modular Utility for STAtistical kNowledge Gathering

NISO National Information Standards Organization

NWP Numerical Weather Prediction

OAI-ORE Open Archives Initiative Object Reuse and Exchange

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting

OAIS Open Archival Information System

OAIS RM Open Archival Information System Reference Model

OGC Open Geospatial Consortium

OLAP Online Analytical Processing

ORFEUS Observatories & Research Facilities for European Seismology

OWL Web Ontology Language

PDF Probabilistic Density Function

PREMIS PREservation Metadata Implementation Strategies

PSD Power Spectral Density

RDA Research Data Alliance

RDF Resource Description Framework

SDI Spatial Data Infrastructure

SG Science Gateway

SHACL Shapes Constraint Language

SKA Square Kilometre Array

SKOS Simple Knowledge Organization System

SLA Service Level Agreement

SOS System of Systems

SUS System Usability Scale

TCS Thematic Core Services

UoD Universe of Discourse

URI Universal Resource Identifier

VRE Virtual Research Environment

VTF Vocabulary Task Force

WCPS Web Coverage Processing Service

WCS Web Coverage Service

Appendix A

EPOS-DCAT-AP class diagram

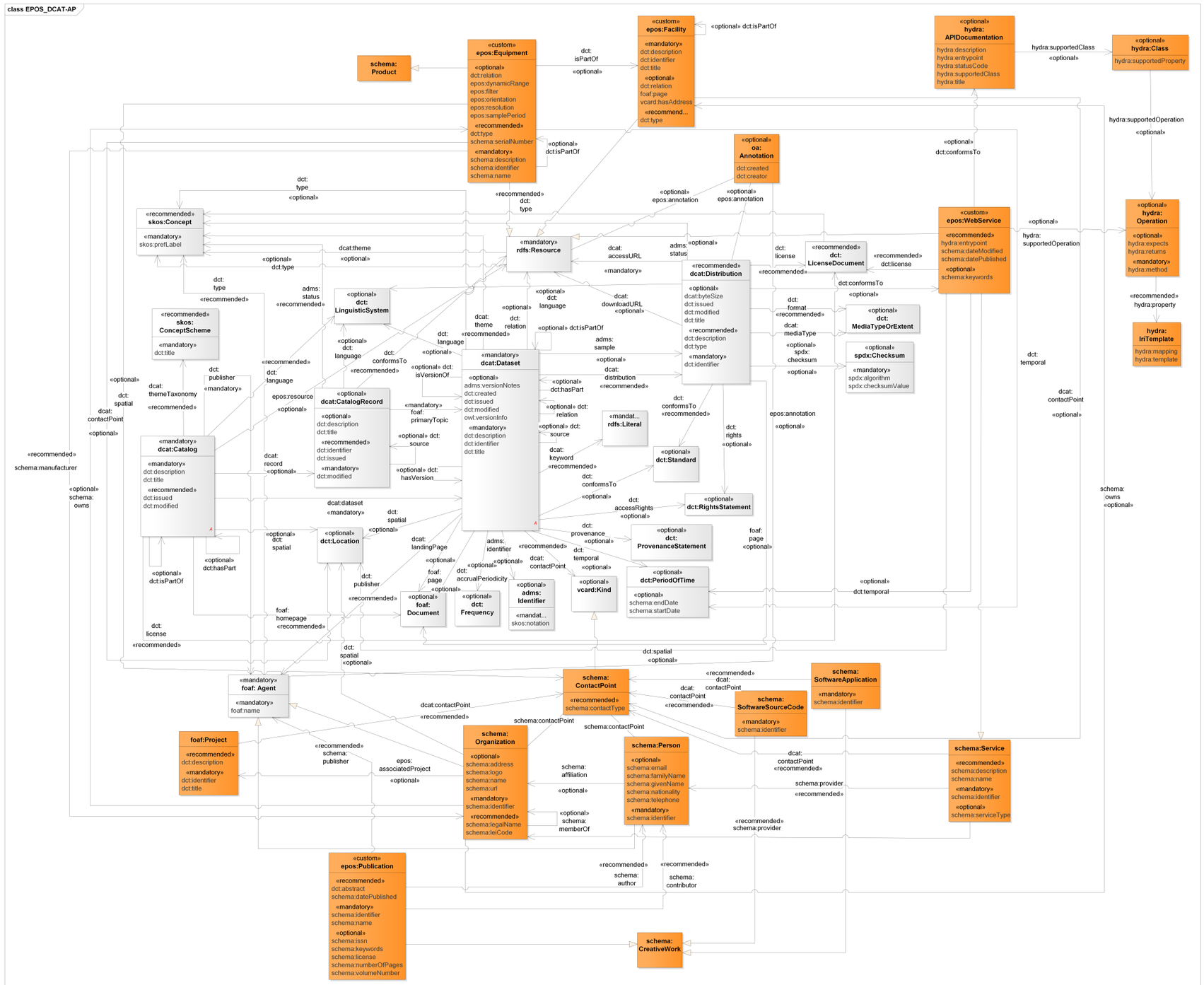


Figure A.1: EPOS-DCAT-AP class diagram – It extends DCAT-AP v1.1 [European Commission, 2015a] (whose classes are depicted in grey) with additional classes (in orange) and relationships. Stereotypes indicate ‘mandatory’, ‘recommended’ and ‘optional’ elements. Similarly, ‘custom’ indicates additional classes defined in the epos namespace.

Appendix B

Definition of EPOS-DCAT-AP

Listing B.1: Definition of the EPOS-DCAT-AP ontology and SHACL shapes graphs

```
@prefix : <http://www.epos-eu.org/epos-dcat-ap#> .
@prefix adms: <http://www.w3.org/ns/adms#> .
@prefix dash: <http://datashapes.org/dash#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix epos: <http://www.epos-eu.org/epos-dcat-ap#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix cnt: <http://www.w3.org/2011/content#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix spdx: <http://spdx.org/rdf/terms#> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
@prefix hydra: <http://www.w3.org/ns/hydra/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://www.epos-eu.org/epos-dcat-ap>
  rdf:type owl:Ontology ;
  dct:abstract "EPOS DCAT Application Profile is an extension of DCAT-AP for solid-Earth sciences
     communities."@en  ;
  dct:creator [  foaf:name "Luca Trani" ; ] ;
  dct:contributor [foaf:name "Rossana Paciello" ; ] ;
  dct:contributor [foaf:name "Manuela Sbarra" ;  ] ;
  dct:contributor [foaf:name "Damian Ulbricht" ; ] ;
  dct:contributor [foaf:name "Sylvain Grellet" ;  ] ;
  dct:contributor [foaf:name "Andy Riddick" ;  ] ;
  dct:contributor [foaf:name "Xiaoliang Wang" ; ] ;
```

```
  dct:created "2018-02-09"^^xsd:date ;
  dct:modified "2019-05-21"^^xsd:date ;
  dct:relation <https://joinup.ec.europa.eu/node/69559> ;
  dct:title "EPOS DCAT Application Profile"@en  ;
  rdfs:comment "This document specifies the set of classes, properties and shapes graphs used in the
     EPOS DCAT Application Profile."@en  ;
  rdfs:label "EPOS-DCAT-Application Profile"@en  ;
  owl:versionInfo "0.15" ;
  foaf:homepage <http://www.epos-eu.org/epos-dcat-ap> ; .
####
# Classes
####
epos:Equipment
  a owl:Class ;
  rdfs:comment "A generic equipment. E.g. a GPS sensor, a seismic station's channel"@en  ;
  rdfs:isDefinedBy <http://www.epos-eu.org/epos-dcat-ap> ;
  rdfs:label "Equipment"@en  ;
  rdfs:subClassOf rdfs:Resource ;
  rdfs:subClassOf schema:Product ;
.
epos:Facility
  a owl:Class ;
  rdfs:comment "A resource representing a Facility. E.g. a laboratory, a seismic station"@en  ;
  rdfs:isDefinedBy <http://www.epos-eu.org/epos-dcat-ap> ;
  rdfs:label "Facility"@en  ;
  rdfs:subClassOf rdfs:Resource ;
.
epos:Publication
  a owl:Class ;
  rdfs:comment "EPOS specialisation of schema:CreativeWork"@en  ;
  rdfs:isDefinedBy <http://www.epos-eu.org/epos-dcat-ap> ;
  rdfs:label "Publication"@en  ;
  rdfs:subClassOf schema:CreativeWork ;
.
epos:WebService
  a owl:Class ;
  rdfs:comment "A service accessible via a Web API"@en  ;
  rdfs:subClassOf schema:Service;
  rdfs:subClassOf rdfs:Resource ;
  rdfs:label "WebService"@en  ;
  rdfs:isDefinedBy <http://www.epos-eu.org/epos-dcat-ap> ;
.
##Alignment schema:Organization and foaf:Organization##
schema:Organization a owl:Class;
  owl:equivalentClass foaf:Organization;
  rdfs:label "Organization"@en;
.
##Alignment schema:Person and foaf:Person##
schema:Person a owl:Class;
  owl:equivalentClass foaf:Person;
```

```
  rdfs:label "Person"@en;
.
##Alignment schema:ContactPoint and vcard:Kind##
schema:ContactPoint a owl:Class;
    owl:equivalentClass vcard:Kind;
    rdfs:label "ContactPoint"@en;
.
####
# Properties
####
epos:resource
  rdf:type owl:ObjectProperty ;
  rdfs:domain dcat:Catalog ;
  rdfs:range rdfs:Resource;
.
##Alignment dcat:contactPoint and schema:contactPoint##
dcat:contactPoint owl:equivalentProperty schema:contactPoint .
##Alignment dcat:keyword and schema:keywords##
dcat:keyword owl:equivalentProperty schema:keywords .
##Extending range and domain of dcat:contactPoint##
dcat:contactPoint
  rdf:type owl:ObjectProperty ;
  rdfs:domain [rdf:type owl:Class ;
  owl:unionOf(dcat:Dataset schema:Organization
  schema:Person epos:Equipment
  schema:SoftwareApplication schema:SoftwareSourceCode
  epos:WebService foaf:Project epos:Facility)];
  rdfs:range [rdf:type owl:Class ;
  owl:unionOf(vcard:Kind schema:ContactPoint)];
.
epos:associatedProject
  rdf:type owl:ObjectProperty;
  rdfs:domain schema:Organization ;
  rdfs:range foaf:Project ;
.
epos:annotation
  rdf:type owl:ObjectProperty ;
  rdfs:domain [rdf:type owl:Class ;
  owl:unionOf(rdfs:Resource foaf:Agent dcat:Distribution)] ;
  rdfs:range oa:Annotation ;
.
##Extending the domain of dcat:theme##
dcat:theme
  rdf:type owl:ObjectProperty ;
  rdfs:comment "The main category/domain of the referred resource."@en  ;
  rdfs:domain [rdf:type owl:Class ;
  owl:unionOf(rdfs:Resource dcat:Dataset)] ;
  rdfs:range skos:Concept ;
.
epos:dynamicRange
```

```
  rdf:type owl:ObjectProperty ;
  rdfs:comment "This property contains the full scale measuring ability, in nT (unit and value)"@en
     ;
  rdfs:domain epos:Equipment ;
  rdfs:range schema:QuantitativeValue ;
.
epos:filter
  rdf:type owl:ObjectProperty ;
  rdfs:comment "This property describes the filter that an instrument might apply to produce data"@en
       ;
  rdfs:domain epos:Equipment ;
  rdfs:range rdfs:Literal ;
.
epos:samplePeriod
  rdf:type owl:ObjectProperty ;
  rdfs:comment "This property contains the sample period in ms"@en  ;
  rdfs:domain epos:Equipment ;
  rdfs:range schema:QuantitativeValue ;
.
epos:orientation
  rdf:type owl:ObjectProperty ;
  rdfs:comment "This property describes how the instrument is oriented."@en  ;
  rdfs:domain epos:Equipment ;
  rdfs:range rdfs:Literal ;
.
##Spatial properties##
epos:northernmostLatitude
  rdf:type owl:DataTypeProperty ;
  rdfs:subPropertyOf geo:lat;
  rdfs:comment "The WGS84 upper bound (max) latitude of a SpatialThing (decimal degrees)"
.
epos:southernmostLatitude
  rdf:type owl:DataTypeProperty ;
  rdfs:subPropertyOf geo:lat;
  rdfs:comment "The WGS84 lower bound (min) latitude of a SpatialThing (decimal degrees)"
.
epos:westernmostLongitude
  rdf:type owl:DataTypeProperty ;
  rdfs:subPropertyOf geo:lon;
  rdfs:comment "The WGS84 lower bound (min) longitude of a SpatialThing (decimal degrees)"
.
epos:easternmostLongitude
  rdf:type owl:DataTypeProperty ;
  rdfs:subPropertyOf geo:lon;
  rdfs:comment "The WGS84 upper bound (max) longitude of a SpatialThing (decimal degrees)"
.
##Extending the range of schema:owns##
schema:owns
 rdf:type owl:ObjectProperty ;
  rdfs:domain [rdf:type owl:Class ;
```

```
  owl:unionOf(schema:Person schema:Organization)];
  rdfs:range [rdf:type owl:Class ;
  owl:unionOf(schema:OwnershipInfo schema:Product epos:Facility)];
.
epos:resolution
  rdf:type owl:ObjectProperty ;
  rdfs:comment "This property contains the resolution in nT"@en  ;
  rdfs:domain epos:Equipment ;
  rdfs:range rdfs:Literal ;
.
####
# SHACL Shapes Graphs
####
epos:DateOrDateTimeDataType
  a sh:NodeShape ;
  sh:description "It checks that a datatype property receives a date or a dateTime literal"@en  ;
  sh:message "The values must be data typed as either xsd:date or xsd:dateTime"@en  ;
  sh:or ( [ sh:datatype xsd:date ; ]
              [ sh:datatype xsd:dateTime ; ] ) ;
.
epos:ContactPointType
  a sh:NodeShape ;
  sh:description "A vcard:Kind or schema:ContactPoint"@en  ;
  sh:message "The values must be data either vcard:Kind or schema:ContactPoint"@en  ;
  sh:or ( [sh:class vcard:Kind ; ]
      [ sh:class schema:ContactPoint ; ]) ;
.
epos:CatalogShape
  a sh:NodeShape ;
  sh:targetClass dcat:Catalog ;
####
# Catalog mandatory properties
####
  sh:property [
      sh:path dct:title ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
  ] ;
  sh:property [
      sh:path dct:description ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
  ] ;
  sh:property [
      sh:path dct:publisher ;
      sh:maxCount 1 ;
      sh:minCount 1 ;
      sh:or (
        [sh:class foaf:Agent ; ]
        [sh:class schema:Organization ; ]) ;
```

```
  ] ;
  sh:property [
      sh:path dcat:dataset ;
      sh:class dcat:Dataset ;
      sh:minCount 1 ;
    ] ;
####
# Catalog recommended properties
####
  sh:property [
      sh:path dct:issued ;
      sh:minCount 1 ;
      sh:message "Release date is recommended. Please fill in a value"@en  ;
      sh:node epos:DateOrDateTimeDataType ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:issued ;
      sh:maxCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path dct:modified ;
      sh:minCount 1 ;
      sh:message "Update/modification date is recommended. Please fill in a value"@en  ;
      sh:node epos:DateOrDateTimeDataType ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:modified ;
      sh:maxCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path dcat:themeTaxonomy ;
      sh:minCount 1 ;
      sh:message "Theme is recommended. Please fill in a value"@en  ;
      sh:datatype xsd:anyURI;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path foaf:homepage ;
      sh:minCount 1 ;
      sh:class foaf:Document ;
      sh:message "Homepage is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path foaf:homepage ;
      sh:maxCount 1 ;
```

```
      sh:class foaf:Document ;
    ] ;
  sh:property [
      sh:path dct:license ;
      sh:minCount 1 ;
      sh:class dct:LicenseDocument ;
      sh:message "License is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:license ;
      sh:maxCount 1 ;
      sh:class dct:LicenseDocument ;
    ] ;
  sh:property [
      sh:path dct:language ;
      sh:minCount 1 ;
      sh:class dct:LinguisticSystem ;
      sh:message "Language is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
####
# Catalog optional properties
####
  sh:property [
      sh:path dct:hasPart ;
      sh:class dcat:Catalog ;
    ] ;
  sh:property [
      sh:path dct:isPartOf ;
      sh:class dcat:Catalog ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:rights ;
      sh:class dct:RightsStatement ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:spatial ;
      sh:class dct:Location ;
    ] ;
  sh:property [
      sh:path dcat:record ;
      sh:class dcat:CatalogRecord ;
    ] ;
  sh:property [
      sh:path epos:resource ;
      sh:class rdfs:Resource ;
    ] ;
```

```
.
epos:CatalogRecordShape
  a sh:NodeShape ;
  sh:targetClass dcat:CatalogRecord;
####
# CatalogRecord mandatory properties
####
  sh:property [
      sh:path foaf:primaryTopic ;
      sh:class dcat:Dataset ;
      sh:maxCount 1 ;
      sh:minCount 1 ;
    ] ;
  sh:property [
      sh:path dct:modified ;
      sh:maxCount 1 ;
      sh:minCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
####
# CatalogRecord recommended properties
####
  sh:property [
      sh:path dct:identifier ;
      sh:minCount 1 ;
      sh:or ( [sh:datatype xsd:string;]
        [sh:datatype xsd:anyURI; ] );
      sh:message "Identifier is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:identifier;
      sh:or ( [sh:datatype xsd:string; ]
        [sh:datatype xsd:anyURI; ] );
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:conformsTo ;
      sh:datatype xsd:anyURI;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:conformsTo ;
      sh:datatype xsd:anyURI;
      sh:minCount 1 ;
      sh:message "ConformsTo is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path adms:status ;
```

```
      sh:maxCount 1 ;
      sh:datatype xsd:anyURI;
    ] ;
  sh:property [
      sh:path adms:status ;
      sh:minCount 1 ;
      sh:datatype xsd:anyURI;
      sh:message "Status is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
  ] ;
  sh:property [
      sh:path dct:issued ;
      sh:maxCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path dct:issued ;
      sh:minCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
      sh:message "Issued is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
####
# CatalogRecord optional properties
####
  sh:property [
      sh:path dct:language ;
      sh:class dct:LinguisticSystem ;
  ] ;
  sh:property [
      sh:path dct:source ;
      sh:class dcat:CatalogRecord ;
      sh:maxCount 1 ;
  ] ;
  sh:property [
      sh:path dct:title ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path dct:description ;
      sh:datatype xsd:string;
    ] ;
.
epos:PersonShape
  a sh:NodeShape ;
  sh:targetClass schema:Person;
####
# Person mandatory properties
####
  sh:property [
```

```
      sh:path schema:identifier ;
      sh:or ([sh:datatype xsd:string;]
        [ sh:datatype xsd:anyURI; ]
        [sh:class schema:PropertyValue;]) ;
      sh:minCount 1 ;
    ] ;
####
# Person optional properties
####
  sh:property [
      sh:path schema:familyName ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
  ] ;
  sh:property [
      sh:path schema:address ;
      sh:or ([sh:datatype xsd:string ;]
        [sh:class schema:PostalAddress; ] ) ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:email ;
      sh:datatype xsd:string;
    ] ;
   sh:property [
      sh:path schema:qualifications ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path schema:telephone ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:url ;
      sh:datatype xsd:anyURI;
    ] ;
  sh:property [
      sh:path schema:affiliation ;
      sh:class schema:Organization;
    ] ;
  sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation;
    ] ;
.
epos:OrganizationShape
```

```
  a sh:NodeShape ;
  sh:targetClass schema:Organization;
####
# Organization mandatory properties
####
  sh:property [
      sh:path schema:identifier ;
      sh:or ([sh:datatype xsd:string; ]
        [sh:datatype xsd:anyURI; ]
        [sh:class schema:PropertyValue; ] ) ;
      sh:minCount 1 ;
    ] ;
####
# Organization recommended properties
####
  sh:property [
      sh:path schema:legalName;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:legalName;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "LegalName is recommended. Please fill in a value"@en ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path schema:leiCode;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:leiCode;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "LeiCode is recommended. Please fill in a value"@en ;
      sh:severity sh:Warning ;
    ] ;
####
# Organization optional properties
####
  sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation;
    ] ;
  sh:property [
      sh:path schema:address;
        sh:or ( [sh:datatype xsd:string ; ]
        [ sh:class schema:PostalAddress; ] ) ;
```

```
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:logo ;
      sh:datatype xsd:anyURI;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:url ;
      sh:datatype xsd:anyURI;
    ] ;
  sh:property [
      sh:path schema:email ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path schema:telephone ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
  ] ;
  sh:property [
      sh:path schema:memberOf;
      sh:class schema:Organization;
   ] ;
  sh:property [
      sh:path schema:owns;
      sh:or ( [sh:class epos:Facility; ]
        [sh:class epos:Equipment; ]);
    ] ;
.
epos:ContactPointShape
  a sh:NodeShape ;
  sh:targetClass schema:ContactPoint;
####
# ContactPoint recommended properties
###
  sh:property [
      sh:path schema:contactType ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "contactType is recommended. Please fill in a value (e.g. legalContact,
     financialContact, scientificContact, manager)"@en  ;
      sh:severity sh:Warning ;
    ] ;
####
# ContactPoint optional properties
####
```

```
  sh:property [
      sh:path schema:name ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
      sh:path schema:email ;
      sh:datatype xsd:string;
    ] ;
  sh:property [
       sh:path schema:availableLanguage ;
       sh:datatype xsd:string;
     ] ;
  sh:property [
      sh:path schema:telephone ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
.
epos:WebServiceShape
  a sh:NodeShape ;
  sh:targetClass epos:WebService;
####
# WebService mandatory properties
####
  sh:property [
      sh:path schema:identifier ;
      sh:or ( [sh:datatype xsd:string; ]
      [ sh:datatype xsd:anyURI; ]
      [ sh:class schema:PropertyValue; ] ) ;
      sh:minCount 1 ;
    ] ;
####
# WebService recommended properties
####
  sh:property [
      sh:path schema:description ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "Description is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path schema:description ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
  sh:property [
     sh:path dcat:theme;
     sh:class skos:Concept ;
     sh:minCount 1 ;
```

```
    sh:message "Theme is recommended. Please fill in a value"@en  ;
    sh:severity sh:Warning ;
   ] ;
 sh:property [
     sh:path schema:name ;
     sh:maxCount 1 ;
     sh:datatype xsd:string;
   ] ;
sh:property [
     sh:path schema:name ;
     sh:minCount 1 ;
     sh:datatype xsd:string;
     sh:message "Name is recommended. Please fill in a value"@en  ;
     sh:severity sh:Warning ;
   ] ;
 sh:property [
     sh:path hydra:entrypoint ;
     sh:maxCount 1 ;
     sh:datatype xsd:anyURI;
   ] ;
 sh:property [
    sh:path hydra:entrypoint ;
    sh:minCount 1 ;
    sh:datatype xsd:anyURI;
    sh:message "Entrypoint is recommended. Please fill in a value"@en  ;
    sh:severity sh:Warning ;
  ] ;
 sh:property [
     sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
     sh:node epos:ContactPointType ;
     sh:minCount 1;
     sh:message "Contact Point is recommended. Please fill in a value"@en  ;
     sh:severity sh:Warning ;
   ] ;
 sh:property [
     sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
     sh:node epos:ContactPointType ;
 ] ;
 sh:property [
     sh:path schema:provider;
     sh:or ( [sh:class schema:Organization ; ]
       [ sh:class schema:Person ; ] );
     sh:maxCount 1;
   ] ;
 sh:property [
     sh:path schema:provider;
     sh:or ( [ sh:class schema:Organization ; ]
       [ sh:class schema:Person ; ] );
     sh:minCount 1;
     sh:message "Provider is recommended. Please fill in a value"@en  ;
```

```
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path schema:datePublished;
      sh:minCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
      sh:message "datePublished is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path schema:datePublished;
      sh:maxCount 1;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path schema:dateModified ;
      sh:minCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
      sh:message "Modified date is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path schema:dateModified;
      sh:node epos:DateOrDateTimeDataType ;
      sh:maxCount 1;
    ] ;
####
# WebService optional properties
####
 sh:property [
      sh:path dct:spatial ;
      sh:class dct:Location;
    ] ;
  sh:property [
      sh:path hydra:supportedOperation ;
      sh:class hydra:Operation ;
    ] ;
  sh:property [
      sh:path dct:conformsTo ;
      sh:class hydra:APIDocumentation ;
    ] ;
  sh:property [
      sh:path schema:keywords;
      sh:datatype xsd:string ;
    ] ;
  sh:property [
      sh:path dct:license ;
      sh:or ( [sh:class dct:LicenseDocument ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:maxCount 1 ;
```

```
  ] ;
  sh:property [
      sh:path dct:temporal ;
      sh:class dct:PeriodOfTime ;
  ] ;
.
epos:OperationShape
a sh:NodeShape ;
sh:targetClass hydra:Operation;
####
# Operation mandatory properties
####
  sh:property [
    sh:path hydra:method ;
    sh:minCount 1 ;
    sh:maxCount 1 ;
    sh:datatype xsd:string ;
 ] ;
####
# Operation  recommended properties
####
  sh:property [
    sh:path hydra:property ;
    sh:class hydra:IriTemplate ;
    sh:minCount 1 ;
    sh:message "IRI template is recommended. Please fill in a value"@en ;
    sh:severity sh:Warning ;
  ] ;
 sh:property [
    sh:path hydra:property ;
    sh:class hydra:IriTemplate ;
    sh:maxCount 1 ;
  ] ;
 sh:property [
    sh:path hydra:returns ;
    sh:datatype xsd:string;
    sh:minCount 1 ;
    sh:message "returns is recommended. Please fill in a value"@en ;
    sh:severity sh:Warning ;
 ] ;
 sh:property [
    sh:path hydra:returns ;
    sh:datatype xsd:string;
    sh:maxCount 1;
 ] ;
####
# Operation optional properties
####
  sh:property [
    sh:path hydra:expects ;
```

```
    sh:class hydra:Class ;
 ] ;
.
epos:ServiceShape
  a sh:NodeShape ;
  sh:targetClass schema:Service;
####
# Service Mandatory properties
####
  sh:property [
      sh:path schema:identifier ;
      sh:or ( [ sh:datatype xsd:string; ]
      [ sh:datatype xsd:anyURI; ]
      [ sh:class schema:PropertyValue; ] ) ;
      sh:minCount 1 ;
    ] ;
####
# Service recommended properties
####
 sh:property [
      sh:path schema:name ;
      sh:maxCount 1 ;
      sh:datatype xsd:string;
 ] ;
 sh:property [
      sh:path schema:name ;
      sh:minCount 1 ;
      sh:datatype xsd:string;
      sh:message "Name is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
];
sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
      sh:minCount 1 ;
      sh:message "Contact point is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
sh:property [
      sh:path schema:description;
      sh:minCount 1 ;
      sh:datatype xsd:string;
      sh:message "Description is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
sh:property [
      sh:path schema:description ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
```

```
sh:property [
    sh:path schema:provider;
    sh:or ( [ sh:class schema:Organization ; ]
      [ sh:class schema:Person ; ] );
    sh:maxCount 1;
  ] ;
sh:property [
    sh:path schema:provider;
    sh:or ( [ sh:class schema:Organization ; ]
      [ sh:class schema:Person ; ] );
    sh:minCount 1;
    sh:message "Provider is recommended. Please fill in a value"@en  ;
    sh:severity sh:Warning ;
  ] ;
####
# Service optional properties
###
sh:property [
      sh:path schema:serviceType ;
      sh:datatype xsd:string;
    ] ;
.
epos:EquipmentShape
  a sh:NodeShape ;
  sh:targetClass epos:Equipment;
####
# Equipment mandatory properties
####
 sh:property [
      sh:path schema:description ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:datatype xsd:string;
 ] ;
 sh:property [
      sh:path schema:identifier ;
      sh:or ( [ sh:datatype xsd:string ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:minCount 1 ;
      sh:maxCount 1 ;
 ] ;
 sh:property [
      sh:path schema:name ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:datatype xsd:string ;
 ] ;
####
# Equipment recommended properties
####
```

```
 sh:property [
      sh:path dct:type ;
      sh:minCount 1 ;
      sh:or ( [ sh:datatype xsd:anyURI ; ]
        [ sh:class skos:Concept ; ] ) ;
      sh:message "Type is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
 ] ;
 sh:property [
      sh:path dct:type ;
      sh:or ( [ sh:datatype xsd:anyURI ; ]
        [ sh:class skos:Concept ; ] ) ;
      sh:maxCount 1 ;
 ] ;
 sh:property [
      sh:path schema:manufacturer ;
      sh:class schema:Organization ;
 ] ;
 sh:property [
      sh:path schema:manufacturer ;
      sh:class schema:Organization ;
      sh:minCount 1;
      sh:message "Manufacturer is recommended. Please fill in a value"@en ;
      sh:severity sh:Warning ;
 ] ;
 sh:property [
      sh:path schema:serialNumber ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
 ] ;
 sh:property [
      sh:path schema:serialNumber ;
      sh:datatype xsd:string ;
      sh:minCount 1 ;
      sh:message "SerialNumber is recommended. Please fill in a value"@en ;
      sh:severity sh:Warning ;
 ] ;
####
# Equipment optional properties
####
 sh:property [
      sh:path dct:isPartOf ;
      sh:or ( [ sh:class epos:Equipment; ]
        [ sh:class epos:Facility; ] );
 ] ;
 sh:property [
      sh:path epos:filter ;
      sh:datatype xsd:string ;
 ] ;
 sh:property [
```

```
      sh:path epos:dynamicRange ;
      sh:class schema:QuantitativeValue ;
 ] ;
 sh:property [
      sh:path epos:orientation ;
      sh:datatype xsd:string ;
 ] ;
 sh:property [
      sh:path epos:resolution ;
      sh:datatype xsd:string ;
 ] ;
 sh:property [
      sh:path epos:samplePeriod ;
      sh:class schema:QuantitativeValue ;
 ] ;
 sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
 ];
 sh:property [
      sh:path dct:spatial ;
      sh:class dct:Location ;
    ] ;
 sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation ;
 ] ;
 sh:property [
       sh:path dct:temporal ;
       sh:class dct:PeriodOfTime ;
 ] ;
sh:property [
      sh:path dct:relation ;
      sh:or ( [ sh:class rdfs:Resource ; ]
        [ sh:class dcat:Dataset ; ]
  [ sh:class epos:WebService ; ] );
] ;
sh:property [
     sh:path dcat:theme ;
     sh:class skos:Concept ;
] ;
.
epos:DatasetShape
  a sh:NodeShape ;
  sh:targetClass dcat:Dataset ;
####
# Dataset mandatory properties
####
 sh:property [
      sh:path dct:description ;
```

```
      sh:minCount 1 ;
      sh:datatype xsd:string ;
    ] ;
 sh:property [
      sh:path dct:identifier ;
      sh:or ( [ sh:datatype xsd:string ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:minCount 1 ;
      sh:maxCount 1 ;
 ] ;
 sh:property [
      sh:path dct:title ;
      sh:minCount 1 ;
      sh:datatype xsd:string ;
 ] ;
####
# Dataset recommended properties
####
  sh:property [
      sh:path dcat:distribution ;
      sh:minCount 1 ;
      sh:class dcat:Distribution;
      sh:message "Distribution is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dcat:distribution ;
      sh:class dcat:Distribution;
    ] ;
  sh:property [
      sh:path dcat:contactPoint ;
      sh:minCount 1 ;
      sh:node epos:ContactPointType ;
      sh:message "Contact point is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
  ] ;
  sh:property [
      sh:path dcat:contactPoint ;
      sh:node epos:ContactPointType ;
    ] ;
  sh:property [
      sh:path dct:publisher ;
      sh:minCount 1 ;
      sh:or ( [ sh:class foaf:Agent ; ]
      [ sh:class schema:Organization ; ] ) ;
      sh:message "Publisher is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
 sh:property [
      sh:path dct:publisher ;
```

```
      sh:maxCount 1 ;
      sh:or ( [ sh:class foaf:Agent ; ]
      [ sh:class schema:Organization ; ] ) ;
    ] ;
 sh:property [
      sh:path dcat:keyword ;
      sh:minCount 1 ;
      sh:datatype xsd:string ;
      sh:message "Keyword is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
 sh:property [
      sh:path dcat:keyword ;
      sh:datatype xsd:string ;
    ] ;
sh:property [
      sh:path dcat:theme ;
      sh:class skos:Concept ;
      sh:minCount 1 ;
      sh:message "Theme is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
] ;
####
# Dataset optional properties
####
  sh:property [
      sh:path dct:created ;
      sh:maxCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path dct:type ;
      sh:datatype xsd:anyURI ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:accessRights ;
      sh:class dct:RightsStatement ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:accrualPeriodicity ;
      #sh:class dct:Frequency ;
      sh:datatype xsd:anyURI ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:conformsTo ;
      sh:class dct:Standard ;
  ] ;
```

```
sh:property [
    sh:path dct:hasVersion ;
    sh:class dcat:Dataset ;
  ] ;
sh:property [
    sh:path dct:isVersionOf ;
    sh:class dcat:Dataset ;
  ] ;
sh:property [
    sh:path dct:issued ;
    sh:maxCount 1 ;
    sh:node epos:DateOrDateTimeDataType ;
  ] ;
sh:property [
    sh:path dct:modified ;
    sh:maxCount 1 ;
    sh:node epos:DateOrDateTimeDataType ;
  ] ;
sh:property [
    sh:path dct:language ;
    sh:class dct:LinguisticSystem ;
  ] ;
sh:property [
    sh:path dct:provenance ;
    sh:class dct:ProvenanceStatement ;
  ] ;
sh:property [
    sh:path dct:relation ;
    sh:class rdfs:Resource ;
  ] ;
sh:property [
    sh:path dct:source ;
    sh:class dcat:Dataset ;
  ] ;
sh:property [
    sh:path dct:spatial ;
    sh:class dct:Location ;
  ] ;
sh:property [
    sh:path dct:temporal ;
    sh:class dct:PeriodOfTime ;
  ] ;
sh:property [
    sh:path owl:versionInfo ;
    sh:maxCount 1 ;
    sh:datatype xsd:string;
 ] ;
sh:property [
    sh:path adms:identifier ;
    sh:class adms:Identifier ;
```

```
    ] ;
  sh:property [
      sh:path adms:sample ;
      sh:class dcat:Distribution ;
    ] ;
  sh:property [
      sh:path dcat:landingPage ;
      sh:class foaf:Document ;
    ] ;
  sh:property [
      sh:path foaf:page ;
      sh:class foaf:Document ;
    ] ;
  sh:property [
      sh:path adms:versionNotes ;
      sh:datatype xsd:string ;
    ] ;
## Support for collections of datasets
  sh:property [
      sh:path dct:isPartOf ;
      sh:class dcat:Dataset ;
    ] ;
  sh:property [
      sh:path dct:hasPart ;
      sh:class dcat:Dataset ;
    ] ;
    sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation ;
    ] ;
.
epos:FacilityShape
  a sh:NodeShape ;
  sh:targetClass epos:Facility ;
####
# Facility mandatory properties
####
  sh:property [
      sh:path dct:description ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:datatype xsd:string ;
    ] ;
  sh:property [
      sh:path dct:identifier ;
      sh:or ( [ sh:datatype xsd:string ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:minCount 1 ;
      sh:maxCount 1 ;
    ] ;
```

```
   sh:property [
      sh:path dct:title ;
      sh:minCount 1 ;
      sh:datatype xsd:string ;
    ] ;
####
# Facility recommended properties
####
  sh:property [
      sh:path dct:type ;
      sh:minCount 1 ;
      sh:or ( [ sh:datatype xsd:anyURI ; ]
        [ sh:class skos:Concept ; ] ) ;
      sh:message "Type is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:type ;
      sh:or ( [ sh:datatype xsd:anyURI ; ]
        [ sh:class skos:Concept ; ] ) ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
     sh:path dcat:theme;
     sh:class skos:Concept ;
     sh:minCount 1 ;
     sh:message "Theme is recommended. Please fill in a value"@en  ;
     sh:severity sh:Warning ;
    ] ;
  ####
  # Facility optional properties
  ####
   sh:property [
      sh:path dct:isPartOf ;
      sh:class epos:Facility ;
   ] ;
   sh:property [
      sh:path vcard:hasAddress ;
      sh:class vcard:Address ;
      sh:maxCount 1 ;
    ] ;
   sh:property [
      sh:path foaf:page ;
      sh:class foaf:Document ;
      sh:maxCount 1 ;
    ] ;
   sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
  ] ;
```

```
  sh:property [
       sh:path dct:relation ;
       sh:or ( [ sh:class rdfs:Resource ; ]
        [ sh:class dcat:Dataset ; ]
  [ sh:class epos:WebService ; ] );
  ] ;
  sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation;
    ] ;
.
epos:DistributionShape
  a sh:NodeShape ;
  sh:targetClass dcat:Distribution;
  ####
  # Distribution mandatory properties
  ####
   sh:property [
      sh:path dct:identifier ;
      sh:or ( [ sh:datatype xsd:string ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:minCount 1 ;
      sh:maxCount 1 ;
    ] ;
   sh:property [
      sh:path dcat:accessURL ;
      sh:or ( [ sh:class rdfs:Resource ; ]
        [ sh:class hydra:Operation ; ]
        [ sh:datatype xsd:anyURI ; ] ) ;
      sh:minCount 1 ;
    ] ;
  ####
  # Distribution recommended properties
  ####
  sh:property [
      sh:path dct:conformsTo ;
      sh:or ( [ sh:class dct:Standard ; ]
        [ sh:class epos:WebService ; ] ) ;
    ] ;
  sh:property [
      sh:path dct:conformsTo ;
      sh:or ( [ sh:class dct:Standard ; ]
        [ sh:class epos:WebService ; ] ) ;
      sh:minCount 1 ;
      sh:message "conformsTo is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:type ;
      sh:minCount 1 ;
```

```
1279      sh:datatype xsd:anyURI;
1280      sh:message "Type is recommended. Please fill in a value"@en  ;
1281      sh:severity sh:Warning ;
1282    ] ;
1283  sh:property [
1284      sh:path dct:type ;
1285      sh:datatype xsd:anyURI ;
1286      sh:maxCount 1 ;
1287    ] ;
1288  sh:property [
1289      sh:path dct:description ;
1290      sh:minCount 1 ;
1291      sh:datatype xsd:string ;
1292      sh:message "Description is recommended. Please fill in a value"@en  ;
1293      sh:severity sh:Warning ;
1294    ] ;
1295  sh:property [
1296      sh:path dct:format ;
1297      sh:or ( [ sh:datatype xsd:string ; ]
1298        [ sh:datatype xsd:anyURI ; ] );
1299      sh:maxCount 1 ;
1300    ] ;
1301   sh:property [
1302      sh:path dct:format ;
1303      sh:or ( [ sh:datatype xsd:string ; ]
1304        [ sh:datatype xsd:anyURI ; ] );
1305      sh:minCount 1 ;
1306      sh:message "Format is recommended. Please fill in a value"@en  ;
1307      sh:severity sh:Warning ;
1308    ] ;
1309  ####
1310  # Distribution optional properties
1311  ####
1312  sh:property [
1313      sh:path dct:issued ;
1314      sh:maxCount 1 ;
1315      sh:node epos:DateOrDateTimeDataType ;
1316    ] ;
1317  sh:property [
1318      sh:path dct:language ;
1319      sh:class dct:LinguisticSystem ;
1320    ] ;
1321  sh:property [
1322      sh:path dct:license ;
1323      sh:or ( [ sh:class dct:LicenseDocument ; ]
1324        [ sh:datatype xsd:anyURI ; ] );
1325      sh:maxCount 1 ;
1326    ] ;
1327  sh:property [
1328      sh:path dct:modified ;
```

```
      sh:maxCount 1 ;
      sh:node epos:DateOrDateTimeDataType ;
    ] ;
  sh:property [
      sh:path dct:rights ;
      sh:class dct:RightsStatement ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dct:title ;
      sh:datatype xsd:string ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path spdx:checksum ;
      sh:class spdx:Checksum ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path adms:status ;
      sh:class skos:Concept ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dcat:byteSize ;
      sh:datatype xsd:decimal ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path dcat:downloadURL ;
      #sh:class rdfs:Resource ;
      sh:datatype xsd:anyURI ;
    ] ;
  sh:property [
      sh:path dcat:mediaType ;
      sh:class dct:MediaTypeOrExtent ;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path foaf:page ;
      sh:class foaf:Document ;
    ] ;
  sh:property [
      sh:path epos:annotation ;
      sh:class oa:Annotation ;
    ] ;
.
epos:ProjectShape
  a sh:NodeShape ;
  sh:targetClass foaf:Project;
```

```
  ####
  # Project mandatory properties
  ####
   sh:property [
      sh:path dct:identifier ;
      sh:or ( [ sh:datatype xsd:string ; ]
        [ sh:datatype xsd:anyURI ; ] );
      sh:minCount 1 ;
    ] ;
    sh:property [
      sh:path dct:title ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:datatype xsd:string ;
    ] ;
  ####
  # Project recommended properties
  ####
  sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
    ] ;
   sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
      sh:minCount 1 ;
      sh:message "Contact point is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
   sh:property [
      sh:path dct:description ;
      sh:minCount 1 ;
      sh:datatype xsd:string;
      sh:message "Description is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
   sh:property [
      sh:path dct:description ;
      sh:datatype xsd:string ;
      sh:maxCount 1 ;
    ] ;
  ####
  # Project optional properties
  ####
  sh:property [
      sh:path foaf:homepage ;
      sh:class foaf:Document ;
  ] ;
.
epos:PublicationShape
```

```
a sh:NodeShape ;
sh:targetClass epos:Publication;
####
# Publication mandatory properties
####
 sh:property [
    sh:path schema:identifier ;
    sh:or ( [ sh:datatype xsd:string ; ]
      [ sh:datatype xsd:anyURI ; ]
      [ sh:class schema:PropertyValue ; ] ) ;
    sh:minCount 1 ;
  ] ;
  sh:property [
    sh:path schema:name ;
    sh:minCount 1 ;
    sh:datatype xsd:string ;
  ] ;
####
# Publication recommended properties
####
sh:property [
    sh:path schema:datePublished ;
    sh:minCount 1 ;
    sh:node epos:DateOrDateTimeDataType ;
    sh:message "Published date is recommended. Please fill in a value"@en  ;
    sh:severity sh:Warning ;
  ] ;
sh:property [
    sh:path schema:datePublished ;
    sh:maxCount 1 ;
    sh:node epos:DateOrDateTimeDataType ;
  ] ;
sh:property [
    sh:path schema:publisher ;
    sh:or ( [ sh:class schema:Person ; ]
    [ sh:class schema:Organization ; ]
    [ sh:class foaf:Agent ; ] ) ;
  sh:maxCount 1;
  ] ;
 sh:property [
    sh:path schema:publisher ;
    sh:or ( [ sh:class schema:Person ; ]
    [ sh:class schema:Organization ; ]
    [ sh:class foaf:Agent ; ] ) ;
    sh:minCount 1 ;
    sh:message "Publisher is recommended. Please fill in a value"@en  ;
    sh:severity sh:Warning ;
  ] ;
  sh:property [
    sh:path dct:abstract ;
```

```
      sh:minCount 1 ;
      sh:datatype xsd:string;
      sh:message "Abstract is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  sh:property [
      sh:path dct:abstract ;
      sh:datatype xsd:string;
      sh:maxCount 1 ;
    ] ;
  sh:property [
      sh:path schema:author ;
      sh:class schema:Person;
    ] ;
  sh:property [
      sh:path schema:author ;
      sh:class schema:Person;
      sh:minCount 1 ;
      sh:message "Author is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
    sh:property [
      sh:path schema:contributor ;
      sh:class schema:Person;
    ] ;
  sh:property [
      sh:path schema:contributor ;
      sh:class schema:Person;
      sh:minCount 1 ;
      sh:message "Contributor is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  ####
  # Publication optional properties
  ####
  sh:property [
      sh:path schema:license ;
      sh:datatype xsd:anyURI;
    ] ;
  sh:property [
      sh:path schema:keywords;
      sh:datatype xsd:string ;
    ] ;
  sh:property [
      sh:path schema:issn ;
      sh:datatype xsd:string ;
    ] ;
  sh:property [
      sh:path schema:numberOfPages ;
      sh:datatype xsd:integer;
```

```
    ] ;
   sh:property [
      sh:path schema:volumeNumber ;
      sh:datatype xsd:string;
    ] ;
.
epos:SoftwareApplicationShape
  a sh:NodeShape ;
  sh:targetClass schema:SoftwareApplication;
  ####
  # SoftwareApplication mandatory properties
  ####
   sh:property [
      sh:path schema:identifier ;
      sh:or ( [ sh:datatype xsd:string; ]
        [ sh:datatype xsd:anyURI; ]
        [sh:class schema:PropertyValue; ] ) ;
      sh:minCount 1 ;
    ] ;
  ####
  # SoftwareApplication recommended properties
  ####
    sh:property [
      sh:path schema:name ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "Name is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
    sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
    ] ;
    sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
      sh:minCount 1 ;
      sh:message "Contact point is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  ####
  # SoftwareApplication optional properties
  ####
    sh:property [
      sh:path schema:description ;
      sh:datatype xsd:string ;
    ] ;
    sh:property [
      sh:path schema:downloadUrl ;
      sh:datatype xsd:anyURI;
```

```
    ] ;
    sh:property [
      sh:path schema:license ;
      sh:datatype xsd:anyURI;
    ] ;
    sh:property [
      sh:path schema:softwareVersion;
      sh:datatype xsd:string ;
    ] ;
.
epos:SoftwareSourceCodeShape
  a sh:NodeShape ;
  sh:targetClass schema:SoftwareSourceCode;
  ####
  # SoftwareSourceCode mandatory properties
  ####
   sh:property [
      sh:path schema:identifier ;
      sh:or ( [ sh:datatype xsd:string; ]
        [ sh:datatype xsd:anyURI; ]
        [sh:class schema:PropertyValue; ] ) ;
      sh:minCount 1 ;
    ] ;
  ####
  # SoftwareSourceCode recommended properties
  ####
    sh:property [
      sh:path schema:name ;
      sh:datatype xsd:string;
      sh:minCount 1 ;
      sh:message "Name is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
    sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
    ] ;
    sh:property [
      sh:path [sh:alternativePath (schema:contactPoint dcat:contactPoint)] ;
      sh:node epos:ContactPointType ;
      sh:minCount 1 ;
      sh:message "Contact point is recommended. Please fill in a value"@en  ;
      sh:severity sh:Warning ;
    ] ;
  ####
  # SoftwareSourceCode optional properties
  ####
    sh:property [
      sh:path schema:description ;
      sh:datatype xsd:string ;
```

```
    ] ;
    sh:property [
      sh:path schema:codeRepository ;
      sh:datatype xsd:anyURI ;
    ] ;
    sh:property [
      sh:path schema:license ;
      sh:datatype xsd:anyURI ;
    ] ;
    sh:property [
      sh:path schema:softwareVersion;
      sh:datatype xsd:string ;
    ] ;
    sh:property [
      sh:path schema:programmingLanguage;
      sh:datatype xsd:string ;
    ] ;
.
```

Bibliography

Ackerman, M. S., Dachtera, J., Pipek, V., and Wulf, V. (2013). Sharing knowledge and expertise: The CSCW view of knowledge management. *Computer Supported Cooperative Work: CSCW: An International Journal*, 22(4-6):531–573.

Ackerman, M. S., Wulf, V., and Pipek, V. (2002). *Sharing Expertise: Beyond Knowledge Management*. MIT Press, Cambridge, MA, USA.

Addair, T. G., Dodge, D. A., Walter, W. R., and Ruppert, S. D. (2014). Large-scale seismic signal analysis with Hadoop. *Computers and Geosciences*, 66:145–154.

Agosti, M., Ferro, N., and Silvello, G. (2016). Digital library interoperability at high level of abstraction. *Future Generation Computer Systems*, 55:129 – 146.

Ahern, T., Casey, R., Barnes, D., Benson, R., Knight, T., and Trabant, C. (2009). *SEED Reference Manual*. IRIS.

Albini, P., Musson, R. M., Capera, A. A. G., Locati, M., Rovida, A., Stucchi, M., and Viganò, D. (2013). Global Historical Earthquake Archive and Catalogue (1000-1903). Technical report, GEM Foundation.

Alemu, G. and Stevens, B. (2015). 1 - Introduction. In Alemu, G. and Stevens, B., editors, *An Emergent Theory of Digital Library Metadata*, pages 1 – 9. Chandos Publishing.

Alemu, G., Stevens, B., and Ross, P. (2012). Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries. *New Library World*, 113(1/2):38–54.

Allemang, D. and Hendler, J. (2008). *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Alobaid, A., Garijo, D., Poveda-Villalón, M., Santana-Perez, I., Fernández-Izquierdo, A., and Corcho, O. (2018). Automating ontology engineering support activities with OnToology. *Journal of Web Semantics*, page 100472.

Angles, R. and Gutierrez, C. (2008). Survey of Graph Database Models. *ACM Comput. Surv.*, 40(1):1:1–1:39.

Apache Software Foundation (2013a). Apache Cassandra. `http://cassandra.apache.org`.

Apache Software Foundation (2013b). Apache CouchDB. `http://couchdb.apache.org`.

Arviset, C. and Gaudet, S. (2010). The IVOA Technical Coordination Group (2010),'IVOA architecture', IVOA Note. *URL: http://www.ivoa.net/documents/Notes/IVOAArchitecture*.

Aryani, A., Burton, A., and Treloar, A. (2016). Research Data Switchboard: Finding Connections to Your Data. `https://figshare.com/articles/Research_Data_Switchboard_Finding_Connections_to_Your_Data/4212261`.

Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., Hausstein, B., Klas, C. P., Zapilko, B., and Kaplun, S. (2018). Data Descriptor: A Research Graph dataset for connecting research data repositories using RD-Switchboard. *Scientific Data*, 5:1–9.

Assante, M., Candela, L., Castelli, D., Coro, G., Lelii, L., and Pagano, P. (2016a). Virtual research environments as-a-service by gCube. *CEUR Workshop Proceedings*, 1871(June):8–10.

Assante, M., Candela, L., Castelli, D., and Tani, A. (2016b). Are Scientific Data Repositories Coping with Research Data Publishing ? *Data Science Journal*, 15:6:1–24.

Atkinson, M. (2018). Pushing the Limits of Data Powered Research. https://doi.org/10.5281/zenodo.1164420.

Atkinson, M., Carpene, M., Casarotti, E., Claus, S., Filgueira, R., Frank, A., Galea, M., Garth, T., Gemund, A., Igel, H., Klampanos, I., Krause, A., Krischer, L., Leong, S. H., Magnoni, F., Matser, J., Michelini, A., Rietbrock, A., Schwichtenberg, H., Spinuso, A., and Vilotte, J.-P. (2015). VERCE Delivers a Productive E-science Environment for Seismology Research. *2015 IEEE 11th International Conference on e-Science (e-Science)*, 00:224–236.

Atkinson, M., Casarotti, E., Eweriing, M., Filgueira, R., Gemünd, A., Klampanos, I., Koukourikos, A., Krause, A., Magnoni, F., Pagani, A., Pagé, C., Rietbrock, A., Spinuso, A., and Wood, C. (2018). D2.1 DARE Architecture & Technical Positioning. Project deliverable, DARE.

Atkinson, M., De Roure, D., Van Hemert, J., and Michaelides, D. (2010). Shaping Ramps for Data-Intensive Research. *UK e-Science All Hands Meeting 2010*, pages 1–3.

Atkinson, M., Gesing, S., Montagnat, J., and Taylor, I. (2017). Scientific workflows: Past, present and future. *Future Generation Computer Systems*, 75(Supplement C):216 – 227.

Atkinson, M. and Parsons, M. (2013). The Digital-Data Challenge. In *The DATA Bonanza*, pages 5–13. John Wiley & Sons, Inc.

Avram, H. D. (1975). *MARC, its history and implications*. Library of Congress,.

Bailo, D., Paciello, R., Rabissoni, R., Sbarra, M., and Vinciarelli, V. (2018). Integration of heterogeneous data, software and services in Solid Earth Sciences: the EPOS system design and roadmap for the building of Integrated Core Services. Technical report, INGV.

Bailo, D., Ulbricht, D., Nayembil, M. L., Trani, L., Spinuso, A., and Jeffery, K. G. (2017). Mapping Solid Earth Data and Research Infrastructures to CERIF. *Procedia Computer Science*, 106:112 – 121.

Bangor, A., Kortum, P., and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies*, 4(3):114–123.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594.

Bannon, L. and Bødker, S. (1997). Constructing Common Information Spaces. *Information Systems Journal*, Proceeding:81–96.

Bannon, L. J. and Kuutti, K. (1996). Shifting perspectives on organizational memory: from storage to active remembering. In *Proceedings of HICSS-29: 29th Hawaii International Conference on System Sciences*, volume 3, pages 156–167 vol.3.

Bannon, L. J. and Schmidt, K. (1989). CSCW: Four characters in search of a context. *Proceedings of the First European Conference on Computer Supported Cooperative Work*, 18(289):358–372.

Barbierato, E., Gribaudo, M., and Iacono, M. (2014). Performance evaluation of NoSQL big-data applications using multi-formalism models. *Future Generation Computer Systems*, 37:345–353.

Baumann, P. (2017). The Datacube Manifesto. `http://www.earthserver.eu/tech/datacube-manifesto`. Last visited on 2018-06-26.

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., Bigagli, L., Boldrini, E., Bruno, R., Calanducci, A., Campalani, P., Clements, O., Dumitru, A., Grant, M., Herzig, P., Kakaletris, G., Laxton, J., Koltsida, P., Lipskoch, K., Mahdiraji, A. R., Mantovani, S., Merticariu, V., Messina, A., Misev, D., Natali, S., Nativi, S., Oosthoek, J., Pappalardo, M., Passmore, J., Rossi, A. P., Rundo, F., Sen, M., Sorbera, V., Sullivan, D., Torrisi, M., Trovato, L., Veratelli, M. G., and Wagner, S. (2016). Big Data Analytics for Earth Sciences: the EarthServer approach. *International Journal of Digital Earth*, 9(1):3–29.

Baumann, P., Rossi, A. P., Bell, B., Clements, O., Evans, B., Hoenig, H., Hogan, P., Kakaletris, G., Koltsida, P., Mantovani, S., Marco Figuera, R., Merticariu, V., Misev,

D., Pham, H. B., Siemen, S., and Wagemann, J. (2018). Fostering Cross-Disciplinary Earth Science Through Datacube Analytics. In Mathieu, P.-P. and Aubrecht, C., editors, *Earth Observation Open Science and Innovation*, pages 91–119. Springer International Publishing, Cham.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., and Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611.

Beltran, A. G., Browning, D., Cox, S., and Winstanley, P. (2018). Data Catalog Vocabulary (DCAT) - revised edition. W3C editor's draft, W3C.

Berners-Lee, T. (1997). Metadata Architecture. `https://www.w3.org/DesignIssues/Metadata.html`. Last visited on 2018-09-26.

Berners-Lee, T. (2006). Linked Data. `https://www.w3.org/DesignIssues/LinkedData.html`. Last visited on 2018-12-20.

Berners-Lee, T. and Connolly, D. (2011). Notation3 N3: A readable RDF syntax. W3C team submission, W3C. https://www.w3.org/TeamSubmission/n3/.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

Bilham, R. and Bendick, R. (2017). A FIVE YEAR FORECAST FOR INCREASED GLOBAL SEISMIC HAZARD. In *Geological Society of America Abstracts with Programs*, volume 49. Invited presentation.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., and Traweek, S. (2015). Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries*, 16(3):207–227.

Brickley, D. and Guha, R. (2014). RDF Schema 1.1. W3C recommendation, W3C.

Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification 0.99. Technical report.

Brooke, J. (1996). SUS: a 'quick and dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation In Industry*, pages 189–194. CRC Press.

Brooke, J. (2013). SUS: A Retrospective. *J. Usability Studies*, 8(2):29–40.

Browning, D. (2018). DXWG DCAT subgroup teleconference 07 November 2018. `https://www.w3.org/2018/11/07-dxwgdcat-minutes#x04`.

Brunsmann, J., Wilkes, W., Schlageter, G., and Hemmje, M. (2012). State-of-the-art of long-term preservation in product lifecycle management. *International Journal on Digital Libraries*, 12(1):27–39.

Bush, V. and Wang, J. (1945). As we may think. *Atlantic Monthly*, 176:101–108.

Byrne, K. (2009). *Populating the semantic web: combining text and relational databases as RDF graphs*. PhD thesis, The University of Edinburgh.

Cabitza, F., Colombo, G., and Simone, C. (2013). Leveraging underspecification in knowledge artifacts to foster collaborative activities in professional communities. *International Journal of Human-Computer Studies*, 71(1):24 – 45. Special Issue on supporting shared representations in collaborative activities.

Cabitza, F., Simone, C., and Sarini, M. (2008). Knowledge Artifacts as Bridges between Theory and Practice: The Clinical Pathway Case. In Ackerman, M., Dieng-Kuntz, R., Simone, C., and Wulf, V., editors, *Knowledge Management In Action: IFIP 20th World Computer Congress, Conference on Knowledge Management in Action, September 7-10, 2008, Milano, Italy*, pages 37–50. Springer US, Boston, MA.

Candela, L., Athanasopoulos, G., Castelli, D., Raheb, K. E., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G., and Ross, S. (2011). The Digital Library Reference Model, D3.2b. Project deliverable, DL.org.

Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva-McPherson, M., Katifori, V., and Schuldt, H. (2007). The DELOS Digital Library Reference Model. Foundations for Digital Libraries (Version 0.96). Technical report, DELOS Network of Excellence on Digital Libraries.

Candela, L., Castelli, D., and Pagano, P. (2013). Virtual Research Environments : An Overview and a Research Agenda. *Data Science Journal*, 12(August):75–81.

Caplan, P. (2017). Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata. Technical report, Library of Congress.

Casey, R., Templeton, M. E., Sharer, G., Keyson, L., Weertman, B. R., and Ahern, T. (2018). Assuring the Quality of IRIS Data with MUSTANG. *Seismological Research Letters*, 89(2A):630.

Cauzzi, C., Sleeman, R., Clinton, J., Ballesta, J. D., Galanis, O., and Kästli, P. (2016). Introducing the European Rapid Raw Strong-Motion Database. *Seismological Research Letters*, 87(4):977–986.

CCSDS (2002). Reference Model for an Open Archival Information System (OAIS), Blue Book, Issue 1. Technical report, CCSDS - Consultative Committee for Space Data Systems.

CCSDS (2012). CCSDS. Reference Model for an Open Archival Information System (OAIS). Magenta Book CCSDS 650.0-M-2, 2012. Technical report, CCSDS - Consultative Committee for Space Data Systems. Also published as ISO 14721:2003.

Chan, L. M. and Zeng, M. L. (2006). Metadata interoperability and standardization – A study of methodology part I: Achieving interoperability at the schema level. *D-Lib Magazine*, 12(6):23–41.

Charalambidis, A., Troumpoukis, A., and Konstantopoulos, S. (2015). SemaGrow: Optimizing Federated SPARQL Queries. In *Proceedings of the 11th International Conference on Semantic Systems*, SEMANTICS '15, pages 121–128, New York, NY, USA. ACM.

Chungoora, N., Young, R. I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N. A., Cutting-Decelle, A.-F., Harding, J. A., and Case, K. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. *Computers in Industry*, 64(4):392 – 401.

Cocco, M. (2018). EPOS IP Management Plan, D1.1. `https://doi.org/10.5281/zenodo.1213698`.

Codd, E. F., Codd, S. B., and Salley, C. T. (1993). Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd and Associates.

Corcho, O., Fernández-López, M., Gómez-Pérez, A., and López-Cima, A. (2005). Building legal ontologies with METHONTOLOGY and WebODE. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3369 LNAI:142–157.

Cox, S. and Little, C. (2017). Time Ontology in OWL. W3C recommendation, W3C.

Cox, S. J. D. (2016). Time ontology extended for non-Gregorian calendar applications. *Semantic Web*, 7(2):201–209.

Cox, S. J. D. and Richard, S. M. (2015). A geologic timescale ontology and service. *Earth Science Informatics*, 8(1):5–19.

Craglia, M. and Nativi, S. (2018). Mind the Gap: Big Data vs. Interoperability and Reproducibility of Science. In Mathieu, P.-P. and Aubrecht, C., editors, *Earth Observation Open Science and Innovation*, pages 121–141. Springer International Publishing, Cham.

Cyganiak, R. and Reynolds, D. (2014). The RDF Data Cube Vocabulary. W3C recommendation, W3C. http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.

DataCite Metadata Working Group (2016). DataCite Metadata Schema for the Publication and Citation of Research Data (Version 2.1). Technical report, DataCite.

DCMI (2000). Memorandum of Understanding between the Dublin Core Metadata Initiative and the IEEE Learning Technology Standards Committee. `http://`

dublincore.org/documents/2000/12/06/dcmi-ieee-mou/. Last visited on 2018-12-10.

DCMI (2012). Dublin Core Metadata Element Set. http://dublincore.org/documents/2012/06/14/dces/. Last visited on 2018-08-07.

DCMI Usage Board (2012). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/. Last visited on 2019-01-31.

Demleitner, M., Greene, G., Le Sidaner, P., and Plante, R. L. (2014). The virtual observatory registry. *Astronomy and Computing*, 7:101–107.

Derriere, S., Gray, A. J. G., Gray, N., Hessman, F. V., Linde, T., Martinez, A. P., and Thomas, B. (2008). Vocabularies in the Virtual Observatory. *Astronomy*, (December):1–16.

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van de Walle, R. (2014). RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of the 7th Workshop on Linked Data on the Web*.

Duggan, J., Elmore, A. J., Stonebraker, M., Balazinska, M., Howe, B., Kepner, J., Madden, S., Maier, D., Mattson, T., and Zdonik, S. (2015). The BigDAWG Polystore System. *ACM Sigmod Record*, 44(3):11–16.

Duval, E., Hodgins, W., Sutton, S., and Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, 8(4):1–15.

EU Parliament (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, 50(L108).

European Commission (2011). Open data An engine for innovation, growth and transparent governance Swedish open data portal – COM(2011) 882 final. http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2011/0882/COM_COM(2011)0882_EN.pdf. Last visited on 2017-11-23.

European Commission (2015a). DCAT Application Profile for data portals in Europe Document Metadata. Technical report, European Commission.

European Commission (2015b). GeoDCAT-AP : A geospatial extension for the DCAT application profile for data portals in Europe. Technical report, European Commission.

European Commission (2016). StatDCAT-AP – DCAT Application Profile for description of statistical datasets. Technical report, European Commission.

European Commission (2017a). COMMISSION IMPLEMENTING DECISION (EU) 2017/1358 of 20 July 2017 on the identification of ICT Technical Specifications for referencing in public procurement. `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017D1358&from=EN`.

European Commission (2017b). European data portal. `www.europeandataportal.eu`. Last visited on 2017-03-21.

European Commission (2018). New European infrastructure to support 'Earth-science' research. `https://ec.europa.eu/info/news/new%2Deuropean%2Dinfrastructure%2Dsupport%2Dearth%2Dscience%2Dresearch%2D2018%2Doct%2D30_en`. Last visited on 2018-11-04.

Facebook Open Source (2018). A query language for your API. `https://graphql.org/`. Last visited on 2018-12-18.

Fan, W., Xu, J., Wu, Y., Yu, W., Jiang, J., Zheng, Z., Zhang, B., Cao, Y., and Tian, C. (2017). Parallelizing Sequential Graph Computations. In *Proceedings of the 2017 ACM International Conference on Management of Data - SIGMOD '17*, volume 10, pages 495–510.

Fecher, B., Friesike, S., and Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, 10(2):1–25.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *AAAI-97 Spring Symposium Series*, SS-97-06:33–40.

Filguiera, R., Klampanos, I., Krause, A., David, M., Moreno, A., and Atkinson, M. (2014). Dispel4Py: A Python Framework for Data-intensive Scientific Computing. In *Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems*, DISCS '14, pages 9–16, Piscataway, NJ, USA. IEEE Press.

Fischer, G. (2001). Communities of Interest: Learning through the Interaction of Multiple Knowledge Systems. *24th IRIS Conference*, 1:1–13.

Galea, M., Rietbrock, A., Spinuso, A., and Trani, L. (2013). Data-Intensive Seismology: Research Horizons. In *The DATA Bonanza*, pages 353–376. John Wiley & Sons, Inc.

Gandon, F. and Schreiber, G. (2014). RDF 1.1 XML Syntax. W3C recommendation, W3C. http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/.

Ganesh Chandra, D. (2015). BASE analysis of NoSQL database. *Future Generation Computer Systems*, 52:13–21.

Gao, S., Sperberg-McQueen, M., and Thompson, H. (2012). W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. W3C recommendation, W3C.

Gartner, R. (2016). What Metadata Is and Why It Matters. In *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*, pages 1–13. Springer International Publishing.

Genova, F., Allen, M. G., Arviset, C., Lawrence, A., Pasian, F., Solano, E., and Wambsganss, J. (2015). Euro-VO—Coordination of virtual observatory activities in Europe. *Astronomy and Computing*, 11:181 – 189. The Virtual Observatory: II.

Genova, F., Arviset, C., Almas, B. M., Bartolo, L., Broeder, D., Law, E., and McMahon, B. (2017). Building a Disciplinary, World-Wide Data Infrastructure. *Data Science Journal*, 16:1–13.

Gesing, S. and Wilkins-Diehr, N. (2015). Science gateway workshops 2014 special issue conference publications. *Concurrency and Computation: Practice and Experience*, 27(16):4247–4251.

Godey, S., Bossu, R., and Guilbert, J. (2013). Improving the mediterranean seismicity picture thanks to international collaborations. *Physics and Chemistry of the Earth*, 63:3–11.

Golbreich, C. and Wallace, E. (2012). OWL 2 Web Ontology Language New Features and Rationale (Second Edition). W3C recommendation, W3C.

Golub, K. (2011). Knowledge organization systems. `http://technicalfoundations.ukoln.ac.uk/print/44.html`. Last visited on 2018-08-13.

Gonçcalves, M. A. (2004). *Streams, Structures, Spaces,Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications*. PhD thesis.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Mining and Knowledge Discovery*, 1(1):29–53.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., and Heber, G. (2005). Scientific data management in the coming decade. *SIGMOD Rec.*, 34(4):34–41.

Greif, I., editor (1988). *Computer-supported Cooperative Work: A Book of Readings*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Gremler, D. D. (2004). The Critical Incident Technique in Service Research. *Journal of Service Research*, 7(1):65–89.

Gruber, T. (2007). Ontology of Folksonomy: A Mash-Up of Apples and Oranges. *International Journal on Semantic Web and Information Systems*, 3(1):1–11.

Gruber, T. R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(5):907 – 928.

Gruninger, M. and Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*, pages 1–10.

Guha, R. V., Brickley, D., and MacBeth, S. (2015). Schema.Org: Evolution of Structured Data on the Web. *Queue*, 13(9):10:10–10:37.

Hameed, A., Preece, A., and Sleeman, D. (2004). Ontology Reconciliation. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 231–250. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hanisch, R. (2014). The Virtual Observatory: I. *Astronomy and Computing*, 7-8:1 – 2. Special Issue on The Virtual Observatory: I.

Hanisch, R., Berriman, G., Lazio, T., Bunn, S. E., Evans, J., McGlynn, T., and Plante, R. (2015). The Virtual Astronomical Observatory: Re-engineering access to astronomical data. *Astronomy and Computing*, 11:190 – 209. The Virtual Observatory: II.

Haslhofer, B. and Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2):1–37.

Heery, R. and Patel, M. (2000). Application profiles : mixing and matching metadata schemas. *AGI - Information Management Consultants*, 10(1):1–10.

Herman, I., Adida, B., Sporny, M., and Birbeck, M. (2015). RDFa 1.1 Primer - Third Edition. W3C note, W3C.

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Number April. The Digital Library Federation Council on Library and Information Resources.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., and Wittenburg, P. (2018). Turning FAIR data into reality. Interim report from the European Commission Expert Group on FAIR data. Technical Report June.

Hou, C.-y., Thompson, C. A., and Palmer, C. L. (2014). Profiling Open Digital Repositories in the Atmospheric and Climate Sciences : An Initial Survey. *77th ASIS&T Annual Meeting*, pages 4–7.

Huysman, M. and Wulf, V. (2004). *Social Capital and Information Technology*. The MIT Press.

Imran, M. and Young, R. (2016). Reference ontologies for interoperability across multiple assembly systems. *International Journal of Production Research*, 54(18):5381–5403.

INSPIRE Maintenance and Implementation Group (MIG) (2016). Technical Guidance for the implementation of INSPIRE Download Services using Web Coverage Services (WCS). Technical report, European Commission. Last visited on 2018-06-25.

Iorio, A. and Caron, B. (2016). PREMIS 3.0 Ontology: Improving Semantic Interoperability of Preservation Metadata. *Proceedings of the 13th International Conference on Digital Preservation (iPRES2016)*, pages 3–7.

Isaac, A. and Summers, E. (2009). SKOS Simple Knowledge Organization System Primer. W3C recommendation, W3C.

Isaac, A., Waites, W., Young, J., and Zeng, M. (2011). Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. W3C incubator group report, W3C.

Ivanova, M., Kargin, Y., Kersten, M., Manegold, S., Zhang, Y., Datcu, M., and Molina, D. E. (2013a). Data vaults: a Database Welcome to Scientific File Repositories. *Proceedings of the 25th International Conference on Scientific and Statistical Database Management - SSDBM*, page 1.

Ivanova, M., Kersten, M., Manegold, S., and Kargin, Y. (2013b). Data Vaults : Database Technology. *Computing in Science & Engineering*, 15(3):32–42.

Jamshidi, M. (2008). System of Systems - Innovations for 21st Century. In *2008 IEEE Region 10 and the Third international Conference on Industrial and Information Systems*, pages 6–7.

Jeffery, K. and Bailo, D. (2014). EPOS: Using Metadata in Geoscience. In Closs, S., Studer, R., Garoufallou, E., and Sicilia, M.-A., editors, *Metadata and Semantics Research*, volume 478 of *Communications in Computer and Information Science*, pages 170–184. Springer International Publishing.

Jeffery, K., Meghini, C., Concordia, C., Patkos, T., Brasse, V., van Ossenbruck, J., Marketakis, Y., Minadakis, N., and Marchetti, E. (2017). A Reference Architecture for Virtual Research Environments. *Proceedings of the 15th International Symposium of Information Science, 13-15 March 2017, Humboldt-Universität zu Berlin*, pages 76–88.

Kim, Y. and Stanton, J. M. (2013). Institutional and Individual Factors Affecting Scientists' Data-Sharing Behaviors: A Multilevel Analysis. *International Review of Research in Open and Distance Learning*, 14(4):90–103.

Knublauch, H., Allemang, D., and Steyskal, S. (2017). SHACL Advanced Features. W3C note, W3C.

Knublauch, H. and Kontokostas, D. (2017). Shapes Constraint Language (SHACL). W3C recommendation, W3C.

Kostkova, P. and Madle, G. (2013). What impact do healthcare digital libraries have? An evaluation of national resource of infection control at the point of care using the Impact-ED framework. *International Journal on Digital Libraries*, 13(2):77–90.

Koymans, M., Fares, M., Trani, L., Quinteros, J., and Nagoe, C. (2018). FAIRYTALE – Towards FAIR Seismological Data Management in the European Integrated Data Archive (EIDA). In *Book of Abstracts of the 36th General Assembly of the European Seismological Commission*. Mistral Service.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J. (2015). ObsPy : a bridge for seismology into the scientific Python ecosystem. *Computational Science & Discovery*, 8(1):1–17.

Krogh, P. G. and Petersen, M. G. (2010). Designing for collective interaction: Toward desirable spaces in homes and libraries. In *From CSCW to Web 2.0: European Developments in Collaborative Design - Selected Papers from COOP 2008*, pages 97–113.

Kunze, J. A., Littman, J., Madden, L., Scancella, J., and Adams, C. (2018). The BagIt File Packaging Format (V1.0). Internet-Draft draft-kunze-bagit-16, Internet Engineering Task Force. Work in Progress.

Lagoze, C. and de Sompel, H. V. (2008). ORE User Guide - Primer. Primer, Open Archives Initiative.

Lagoze, C. and de Sompel, H. V. (2017). Open archives initiative resourcesync framework specification. Specification, Open Archives Initiative.

Lannom, L. and Wittenburg, P. (2016). Global Digital Object Cloud. `http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3`.

Lanthaler, M. (2013). Creating 3rd Generation Web APIs with Hydra. *Proceedings of the 22nd International World Wide Web Conference (WWW2013)*, pages 35–37.

Lanthaler, M. (2014). *Third Generation Web APIs - Bridging the Gap between REST and Linked Data*. PhD thesis.

Lanthaler, M. (2018). Hydra Core Vocabulary. Unofficial draft, W3C.

Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press.

Lavoie, B. (2014). Information System (OAIS) Reference Model : Introductory Guide. Technical report, Digital Preservation Coalition.

Le-Phuoc, D., Nguyen Mau Quoc, H., Ngo Quoc, H., Tran Nhat, T., and Hauswirth, M. (2016). The Graph of Things: A step towards the Live Knowledge Graph of connected things. *Journal of Web Semantics*, 37-38:25–35.

Lebo, T., Sahoo, S., and McGuinness, D. (2013). PROV-o: The PROV ontology. W3C recommendation, W3C. http://www.w3.org/TR/2013/REC-prov-o-20130430/.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Lei Zeng, M. (2008). Knowledge Organization Systems. *Knowledge organization*, 35(2-3):160–182.

Lei Zeng, M. and Mayr, P. (2018). Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review. *International Journal on Digital Libraries*, pages 1–22.

Leidig, J. P. and Fox, E. a. (2014). Intelligent digital libraries and tailored services. *Journal of Intelligent Information Systems*, 43(3):463–480.

Levin, B. W., Sasorova, E. V., Steblov, G. M., Domanski, A. V., Prytkov, A. S., and Tsyba, E. N. (2017). Variations of the Earth's rotation rate and cyclic processes in geodynamics. *Geodesy and Geodynamics*, 8(3):206–212.

Li, Z., Yang, M. c., and Ramani, K. (2009). A Methodology for Engineering Ontology Acquisition and Validation. *Artif. Intell. Eng. Des. Anal. Manuf.*, 23(1):37–51.

Lins, L., Klosowski, J. T., and Scheidegger, C. (2013). Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465.

Loshin, D. (2013a). Chapter 10 - Using Graph Analytics for Big Data. In Loshin, D., editor, *Big Data Analytics*, pages 91 – 103. Morgan Kaufmann, Boston.

Loshin, D. (2013b). Chapter 9 - NoSQL Data Management for Big Data. In Loshin, D., editor, *Big Data Analytics*, pages 83 – 90. Morgan Kaufmann, Boston.

Lourenço, J. R., Cabral, B., Carreiro, P., Vieira, M., and Bernardino, J. (2015). Choosing the right NoSQL database for the job: a quality attribute evaluation. *Journal of Big Data*, 2(1):18.

Lubich, H. (1995a). *Towards a CSCW Framework for Scientific Cooperation in Europe*. Springer Berlin Heidelberg.

Lubich, H. P. (1995b). Foundations of scientific cooperation. In *Towards a CSCW Framework for Scientific Cooperation in Europe*, pages 60–73. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ma, X., Fu, L., West, P., and Fox, P. (2018). Ontology Usability Scale: Context-aware Metrics for the Effectiveness, Efficiency and Satisfaction of Ontology Uses. *Data Science Journal*, 17(1995):1–11.

Maali, F. and Erickson, J. (2014). Data Catalog Vocabulary (DCAT). W3C recommendation, W3C.

Magagna, B., Goldfarb, D., Martin, P., Toussaint, F., Kindermann, S., Atkinson, M., Jeffery, K., Fiebig, M., de la Hidalga, A. N., and Spinuso, A. (2018). D8.5 Data provenance and tracing for environmental sciences: system design. Project deliverable, ENVRIplus.

Maidment, D., Domenico, B., Gemmell, A., Lehnert, K., Tarboton, D., and Zaslavsky, I. (2011). The Open Geospatial Consortium and EarthCube. Technical report, EarthCube Technology.

Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., and Pagano, P. (2014). The D-NET software toolkit. *Program*, 48(4):322–354.

Manola, F. and Miller, E. (2004). RDF Primer. W3C recommendation, W3C.

Marcial, L. H. and Hemminger, B. M. (2013). Scientific Data Repositories on the Web: An Initial Survey. *International Review of Research in Open and Distance Learning*, 14(4):90–103.

Marshal, M. (2011). *Ontology spectrum for geological data interoperability*. PhD thesis, University of Twente.

McDonough, J. P. (2011). Packaging videogames for long-term preservation: Integrating FRBR and the OAIS reference model. *Journal of the American Society for Information Science and Technology*, 62(1):171–184.

McGibbney, L. (2018). Semantic Web for Earth and Environmental Terminology(SWEET) 2018:Status, Future Development and Community Building. `https://esip.figshare.com/articles/McGibbneyL_sweet_20180110_pdf/5782122`.

McGuinness, D. and van Harmelen, F. (2004). OWL Web Ontology Language Overview. W3C recommendation, W3C.

McMeekin, S. M. (2011). With a Little Help from OAIS: Starting down the Digital Curation Path. *Journal of the Society of Archivists*, 32(2):241–253.

McNamara, D. E. and Boaz, R. I. (2006). PQLX: A Software Tool to Evaluate Seismic Station Performance. *AGU Fall Meeting Abstracts*.

Members, PALAEOSENS Project (2012). Making sense of palaeoclimate sensitivity. *Nature*, 491:683. Perspective.

Miles, A. and Bechhofer, S. (2009a). SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant. W3C recommendation, W3C.

Miles, A. and Bechhofer, S. (2009b). SKOS Simple Knowledge Organization System Reference. W3C recommendation, W3C.

Minadakis, N., Marketakis, Y., Kondylakis, H., Flouris, G., Theodoridou, M., Doerr, M., and de Jong, G. (2015). X3ML Framework: An effective suite for supporting data mappings. *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL2015)*, pages 1–12.

MonetDB BV (2013). MonetDB. `www.monetdb.org`.

MongoDB, Inc. (2016). MongoDB. `www.mongodb.com`.

Musen, M. A. (2015). The Protégé Project: A Look Back and a Look Forward. *AI Matters*, 1(4):4–12.

National Research Council (2001). *Astronomy and Astrophysics in the New Millennium*. The National Academies Press, Washington, DC.

Nativi, S., Mazzetti, P., and Craglia, M. (2017). A view-based model of data-cube to support big earth data systems interoperability. *Big Earth Data*, 4471(December):1–25.

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., and Ochiai, O. (2015). Big Data challenges in building the Global Earth Observation System of Systems. *Environmental Modelling and Software*, 68:1–26.

Neo4J (2016). openCypher. http://www.opencypher.org. Last visited on 2018-11-27.

Nevile, C. M., Brickley, D., and Hickson, I. (2018). HTML Microdata. W3C working draft, W3C.

Nilsson, M. (2010). *From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization*. PhD thesis, KTH, Stockholm.

Nilsson, M. and Johnston, P. (2006). Towards an interoperability framework for metadata standards. *International Conference on Dublin Core and Metadata Applications*, (March 2005).

NISO (2004). Understanding Metadata. *National Information Standards*, (MD:NISO Press):20.

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory*, page 25.

Ontology Engineering Group (2019). Linked Open Terms (LOT) Methodology. https://doi.org/10.5281/zenodo.2539305.

Open Archives Initiative (2002). OAI-PMH. www.openarchives.org/pmh/. Last visited on 2018-06-21.

Open Knowledge Foundation (2013). CKAN. https://ckan.org/. Last visited on 2018-08-21.

Open Knowledge International (2017). Swedish open data portal. www.opengov.se. Last visited on 2017-03-21.

Open Source Geospatial Foundation. (2004). GeoNetwork open source. https://geonetwork-opensource.org/. Last visited on 2018-08-21.

Paciello, R., Trani, L., and Bailo, D. (2019). EPOS-DCAT-AP: an extension of the DCAT Application Profile for Research Infrastructures in the solid-Earth domain. Draft specification, EPOS.

Pierce, M. E., Miller, M. A., Brookes, E. H., Wong, M., Afgan, E., Liu, Y., Gesing, S., Dahan, M., Marru, S., and Walker, T. (2018). Towards a Science Gateway Reference Architecture. *Proceedings of the 10th International Workshop on Science Gateways (IWSG 2018),13-15 June 2018. Edinburgh, Scotland.*

Pons, X. and Masó, J. (2016). A comprehensive open package format for preservation and distribution of geospatial data and metadata. *Computers and Geosciences*, 97(September 2015):89–97.

Powell, J. and Hopkins, M. (2015a). Graph databases and how to use them. In *A Librarian's Guide to Graphs, Data and the Semantic Web*, pages 197–207. Chandos Publishing.

Powell, J. and Hopkins, M. (2015b). Ontologies. In *A Librarian's Guide to Graphs, Data and the Semantic Web*, pages 31–43. Chandos Publishing.

Prabhune, A., Stotzka, R., Sakharkar, V., Hesser, J., and Gertz, M. (2017). MetaStore: an adaptive metadata management framework for heterogeneous metadata models. *Distributed and Parallel Databases*, 36(1):1–42.

PREMIS Working Group (2005). Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Technical report, CLC and RLG.

Prud'hommeaux, E. and Buil-Aranda, C. (2013). SPARQL 1.1 Federated Query. W3C recommendation, W3C.

Prud'hommeaux, E. and Carothers, G. (2014). RDF 1.1 Turtle. W3C recommendation, W3C. http://www.w3.org/TR/2014/REC-turtle-20140225/.

Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. W3C recommendation, W3C.

PwC EU Services (2013). Process and methodology for developing semantic agreements. Technical report, European Commission.

PwC EU Services (2017). Analysis of the DCAT-AP extensions. Technical report, European Commission.

Quimbert, E., Jeffery, K. G., Martens, C., Boulanger, D., Carval, T., Hellström, M., Lankreijer, H., Peterseil, J., Pichot, C., and Zhao, Z. (2018). D8.4 Interoperable cataloguing and metadata harmonisation for environmental RIs: prototype. Project deliverable, ENVRIplus.

Rashid, M. R. A., Rizzo, G., Torchiano, M., Mihindukulasooriya, N., Corcho, O., and García-Castro, R. (2018). Completeness and consistency analysis for evolving knowledge bases. *Journal of Web Semantics*.

Riccardo Rabissoni (2018). EPOS Web Metadata Editor. `http://epos.cineca.it/apache/mde/public/index.php`. Last visited on 2018-11-04.

Riley, J. (2017). *Understanding Metadata What Is Metadata, And What Is It For?* National Information Standards Organization.

Ringler, A. T., Hagerty, M. T., Holland, J., Gonzales, A., Gee, L. S., Edwards, J. D., Wilson, D., and Baker, A. M. (2015). The data quality analyzer: A quality control program for seismic data. *Computers and Geosciences*, 76:96–111.

Sanderson, R., Ciccarese, P., and Young, B. (2017). Web Annotation Vocabulary. W3C recommendation, W3C.

Santoro, M., Nativi, S., and Mazzetti, P. (2016). Contributing to the GEO Model Web implementation: A brokering service for business processes. *Environmental Modelling and Software*, 84:18–34.

Sauro, J. (2018). 5 Ways to Interpret a SUS Score. `https://measuringu.com/interpret-sus-score/`. Last visited on 2018-11-04.

Sauro, J. and Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, (August):2215.

Schmidt, K. and Bannon, L. (1992). Taking CSCW seriously - Supporting articulation work. *Computer Supported Cooperative Work*, 1(1-2):7–40.

Schuh, H., Anderson, J., Beyerle, G., Dick, G., Flechtner, F., Förste, C., Ge, M., Glaser, S., Heinkelmann, R., König, R., Männel, B., Michaelis, I., Quinteros, J., Ramatschi,

M., Rauberg, J., Rother, M., Schmidt, T., Stolle, C., and Wickert, J. (2018). Big Data in Geodäsie, Seismologie und Geomagnetismus. *System Erde; 8*.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101.

Shirky and Clay (2005). Ontology is overrated: categories, links, and tags. `www.shirky.com/writings/ontology_overrated.html`. Last visited on 2017-06-02.

Sleeman, R. (2014a). Automatic data QC and distribution statistics for data providers, D2.3. Project deliverable, NERA.

Sleeman, R. (2014b). Data quality improvement statistics, D2.4. Project deliverable, NERA.

Smith, M., Welty, C., and McGuinness, D. (2004). OWL Web Ontology Language Guide. W3C recommendation, W3C.

Speicher, S., Arwe, J., and Malhotra, A. (2015). Linked Data Platform 1.0. W3C recommendation, W3C.

Sporny, M., Kellogg, G., and Lanthaler, M. (2014). JSON-LD 1.0. W3C recommendation, W3C.

Star, S. L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, 35(5):601–617.

Star, S. L. and Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3):387–420.

Steyskal, S. and Coyle, K. (2017). SHACL Use Cases and Requirements. W3C note, W3C.

Stonebraker, M., Madden, S., Abadi, D. J., Harizopoulos, S., Hachem, N., and Helland, P. (2007). The End of an Architectural Era: (It's Time for a Complete Rewrite). In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 1150–1160. VLDB Endowment.

Strauss, A. (1985). Work and the Division of Labor. *Sociological Quarterly*, 26(1):1–19.

Strobl, P., Baumann, P., Lewis, A., Szantoi, Z., Killough, B., Purss, M., Craglia, M., Nativi, S., Held, A., and Dhu, T. (2017). The Six faces of the Data Cube. *Proceedings of the 2017 conference on Big Data from Space. BIDS' 2017*, (November):17–20.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2015). The NeOn Methodology Framework: A Scenario-Based Methodology for Ontology Development. *Applied Ontology*, 10(2):107–145.

Szejka, A. L. and Junior, O. C. (2017). The Application of Reference Ontologies for Semantic Interoperability in an Integrated Product Development Process in Smart Factories. *Procedia Manufacturing*, 11:1375 – 1384. 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.

The ObsPy Development Team (2016). Obspy 1.0.1. `https://doi.org/10.5281/zenodo.48254`.

Theodoridou, M., Ivanovic, D., Martin, P., Remy, L., and Muckensturm, M. (2019). X3ML mappings from common metadata schemes to CERIF RDF. `https://doi.org/10.5281/zenodo.2548732`.

Tiropanis, T., Hall, W., Hendler, J., and de Larrinaga, C. (2014). The Web Observatory: A Middle Layer for Broad Data. *Big Data*, 2(3):129–133.

Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., and Hendler, J. (2013). The web science observatory. *IEEE Intelligent Systems*, 28(2):100–104.

TopQuadrant (2018). GraphQL Schemas to RDF/SHACL. `https://www.topquadrant.com/graphql/graphql-shacl.html`. Last visited on 2018-12-18.

Trani, L., Atkinson, M., Bailo, D., Paciello, R., and Filgueira, R. (2018a). Establishing Core Concepts for Information-Powered Collaborations. *Future Generation Computer Systems*, 89:421 – 437.

Trani, L., Koymans, M., Atkinson, M., Sleeman, R., and Filgueira, R. (2017). WFCatalog: A catalogue for seismological waveform data. *Computers & Geosciences*, 106:101 – 108.

Trani, L., Paciello, R., Bailo, D., and Vinciarelli, V. (2018b). EPOS-DCAT-AP: a DCAT Application Profile for solid-Earth sciences. In *2018 Fall Meeting AGU*. Abstract IN31B-33.

Trani, L., Paciello, R., Sbarra, M., Ulbricht, D., and the EPOS IT Team (2018c). Representing Core Concepts for solid-Earth sciences with DCAT – the EPOS-DCAT Application Profile. In *Geophysical Research Abstracts*, volume 20.

Trani, L., Sleeman, R., Koymans, M., and the EIDA team (2016). WFCatalog Web Service Specification. `https://www.orfeus-eu.org/documents/WFCatalog_Specification-v0.22.pdf`.

Trani, L. and the EPOS-ORFEUS-CC Team (2018). The EPOS - ORFEUS Competence Centre in EOSC-hub. *The EPOS newsletter*, 03(03).

US Department of Health and Human Services (2013). System Usability Scale (SUS). `https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html`. Last visited on 2018-11-04.

Uyar, A. and Aliyu, F. M. (2015). Evaluating search features of Google Knowledge Graph and Bing Satori. *Online Information Review*, 39(2):197–213.

van der Graaf, M. (2009). *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*. Amsterdam University Press.

van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034.

Vardigan, M. and Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. *Archival Science*, 7(1):73–87.

Vecsey, L. (2018). Experience with the AASN data downloading, implementation of additional data quality control and application of a new arrival-time picker. http://www.alparray.ethz.ch/export/sites/alparray/.galleries/dwn-experiments/2018_AlpArray_meeting_Zurich_LVecsey.pdf. Last visited on 2018-10-28.

Veltman, K. H. (2001). Syntactic and semantic interoperability: New approaches to knowledge and the semantic web. *New Review of Information Networking*, 7(1):159–183.

W3C-DXWG (2018). DXWG Wiki. https://www.w3.org/2017/dxwg/wiki/Main_Page. Last visited on 2018-10-21.

Wal, V. and Thomas (2007). Folksonomy: coinage and definition. http://vanderwal.net/folksonomy.html. Last visited on 2017-06-02.

Wenger, E. (1998). *Communities of practice : learning, meaning, and identity*. Learning in doing. Cambridge University Press, Cambridge, [England] ; New York, N.Y.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., a.C 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. a., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Zeginis, D., Hasnain, A., Loutas, N., Deus, H. F., Fox, R., and Tarabanis, K. (2014). A Collaborative Methodology for Developing a Semantic Model for Interlinking Cancer Chemoprevention Linked-data Sources. *Semant. web*, 5(2):127–142.

Zhang, Z. (2016). *Supporting Information Sharing in Emergency Medical Settings*. PhD thesis, Drexel University. http://hdl.handle.net/1860/idea:7121.